Meta Analysis

A Guide to Calibrating and Combining Statistical Evidence

Elena Kulinskaya
Statistical Advisory Service, Imperial College, London, UK

Stephan Morgenthaler
Institut de Mathématiques, École Polytechnique Fédérale de Lausanne, Switzerland

Robert G. Staudte
Department of Mathematícs and Statistics, La Trobe University, Australia

John Wiley & Sons, Ltd

Meta Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors
David J. Balding, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Sanford Weisberg

Editors Emeriti
Vic Barnett, J. Stuart Hunter, David G. Kendall, Jozef L. Teugels

A complete list of the titles in this series appears at the end of this volume.

Meta Analysis

A Guide to Calibrating and Combining Statistical Evidence

Elena Kulinskaya
Statistical Advisory Service, Imperial College, London, UK

Stephan Morgenthaler
Institut de Mathématiques, École Polytechnique Fédérale de Lausanne, Switzerland

Robert G. Staudte
Department of Mathematícs and Statistics, La Trobe University, Australia

John Wiley & Sons, Ltd

Copyright © 2008 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, ONT, L5R 4J3

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Kulinskaya, Elena.
Meta analysis : a guide to calibrating and combining statistical evidence / Elena Kulinskaya, Stephan Morgenthaler, Robert G. Staudte.
p. ; cm. – (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 978-0-470-02864-3 (pbk. : alk paper)
1. Meta-analysis. I. Morgenthaler, Stephan. II. Staudte, Robert G. III. Title.
IV. Series.
[DNLM: 1. Meta-Analysis. WA 950 K96m 2008]
R853.M48K85 2008
610.72′7–dc22
2007045548

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-470-02864-3 (PB)

Typeset in 10/12pt Times by Thomson Digital Noida
Printed and bound in Great Britain by TJ International, Padstow, Cornwall

To Yuri, Phaik and Heather

Contents

Preface xiii

Part I The Methods 1

1 What can the reader expect from this book? 3

1.1 A calibration scale for evidence . . . 4

1.1.1 T-values and p-values . . . 4

1.1.2 How generally applicable is the calibration scale? . . . 6

1.1.3 Combining evidence . . . 7

1.2 The efficacy of glass ionomer versus resin sealants for prevention of caries . . . 8

1.2.1 The data . . . 8

1.2.2 Analysis for individual studies . . . 9

1.2.3 Combining the evidence: fixed effects model . . . 10

1.2.4 Combining the evidence: random effects model . . . 10

1.3 Measures of effect size for two populations . . . 11

1.4 Summary . . . 13

2 Independent measurements with known precision 15

2.1 Evidence for one-sided alternatives . . . 15

2.2 Evidence for two-sided alternatives . . . 18

2.3 Examples . . . 19

2.3.1 Filling containers . . . 19

2.3.2 Stability of blood samples . . . 20

2.3.3 Blood alcohol testing . . . 20

3 Independent measurements with unknown precision 23

3.1 Effects and standardized effects . . . 23

3.2 Paired comparisons . . . 26

3.3 Examples . . . 27

3.3.1 Daily energy intake compared to a fixed level . . . 27

3.3.2 Darwin's data on *Zea mays* . . . 28

4 Comparing treatment to control **31**
4.1 Equal unknown precision 31
4.2 Differing unknown precision. 33
4.3 Examples 35
4.3.1 Drop in systolic blood pressure 35
4.3.2 Effect of psychotherapy on hospital length of stay 37

5 Comparing *K* treatments **39**
5.1 Methodology 39
5.2 Examples 42
5.2.1 Characteristics of antibiotics 42
5.2.2 Red cell folate levels. 43

6 Evaluating risks **47**
6.1 Methodology 47
6.2 Examples 49
6.2.1 Ultrasound and left-handedness 49
6.2.2 Treatment of recurrent urinary tract infections 49

7 Comparing risks **51**
7.1 Methodology 51
7.2 Examples 54
7.2.1 Treatment of recurrent urinary tract infections 54
7.2.2 Diuretics in pregnancy and risk of pre-eclamsia 54

8 Evaluating Poisson rates **57**
8.1 Methodology 57
8.2 Example. 60
8.2.1 Deaths by horse-kicks 60

9 Comparing Poisson rates **63**
9.1 Methodology 64
9.1.1 Unconditional evidence 64
9.1.2 Conditional evidence. 65
9.2 Example. 67
9.2.1 Vaccination for the prevention of tuberculosis 67

10 Goodness-of-fit testing **71**
10.1 Methodology 71
10.2 Example. 74
10.2.1 Bellbirds arriving to feed nestlings 74

11 Evidence for heterogeneity of effects and transformed effects **77**
11.1 Methodology 77
11.1.1 Fixed effects 77
11.1.2 Random effects. 80

11.2 Examples 81
11.2.1 Deaths by horse-kicks 81
11.2.2 Drop in systolic blood pressure 82
11.2.3 Effect of psychotherapy on hospital length of stay 83
11.2.4 Diuretics in pregnancy and risk of pre-eclamsia 84

12 Combining evidence: fixed standardized effects model 85
12.1 Methodology 86
12.2 Examples 87
12.2.1 Deaths by horse-kicks 87
12.2.2 Drop in systolic blood pressure 88

13 Combining evidence: random standardized effects model 91
13.1 Methodology 91
13.2 Example 94
13.2.1 Diuretics in pregnancy and risk of pre-eclamsia 94

14 Meta-regression 95
14.1 Methodology 95
14.2 Commonly encountered situations 98
14.2.1 Standardized difference of means 98
14.2.2 Difference in risk (two binomial proportions) 99
14.2.3 Log relative risk (two Poisson rates) 99
14.3 Examples 100
14.3.1 Effect of open education on student creativity 100
14.3.2 Vaccination for the prevention of tuberculosis 101

15 Accounting for publication bias 105
15.1 The downside of publishing 105
15.2 Examples 107
15.2.1 Environmental tobacco smoke 107
15.2.2 Depression prevention programs 109

Part II The Theory 111

16 Calibrating evidence in a test 113
16.1 Evidence for one-sided alternatives 114
16.1.1 Desirable properties of one-sided evidence 115
16.1.2 Connection of evidence to *p*-values 115
16.1.3 Why the *p*-value is hard to understand 116
16.2 Random *p*-value behavior 118
16.2.1 Properties of the random *p*-value distribution 118
16.2.2 Important consequences for interpreting *p*-values 119
16.3 Publication bias 119
16.4 Comparison with a Bayesian calibration 121
16.5 Summary 123

17 The basics of variance stabilizing transformations 125
17.1 Standardizing the sample mean . . . 125
17.2 Variance stabilizing transformations . . . 126
17.2.1 Background material . . . 126
17.2.2 The Key Inferential Function . . . 127
17.3 Poisson model example . . . 128
17.3.1 Example of counts data . . . 129
17.3.2 A simple *vst* for the Poisson model . . . 129
17.3.3 A better *vst* for the Poisson model . . . 132
17.3.4 Achieving a desired expected evidence . . . 132
17.3.5 Confidence intervals . . . 132
17.3.6 Simulation study of coverage probabilities . . . 134
17.4 Two-sided evidence from one-sided evidence . . . 134
17.4.1 A *vst* based on the chi-squared statistic . . . 135
17.4.2 A *vst* based on doubling the *p*-value . . . 137
17.5 Summary . . . 138

18 One-sample binomial tests 139
18.1 Variance stabilizing the risk estimator . . . 139
18.2 Confidence intervals for *p* . . . 140
18.3 Relative risk and odds ratio . . . 142
18.3.1 One-sample relative risk . . . 143
18.3.2 One-sample odds ratio . . . 144
18.4 Confidence intervals for small risks *p* . . . 145
18.4.1 Comparing intervals based on the log and arcsine transformations . . . 145
18.4.2 Confidence intervals for small *p* based on the Poisson approximation to the binomial . . . 146
18.5 Summary . . . 147

19 Two-sample binomial tests 149
19.1 Evidence for a positive effect . . . 149
19.1.1 Variance stabilizing the risk difference . . . 149
19.1.2 Simulation studies . . . 151
19.1.3 Choosing sample sizes to achieve desired expected evidence . . . 151
19.1.4 Implications for the relative risk and odds ratio . . . 153
19.2 Confidence intervals for effect sizes . . . 153
19.3 Estimating the risk difference . . . 155
19.4 Relative risk and odds ratio . . . 155
19.4.1 Two-sample relative risk . . . 155
19.4.2 Two-sample odds ratio . . . 157
19.4.3 New confidence intervals for the RR and OR . . . 157
19.5 Recurrent urinary tract infections . . . 157
19.6 Summary . . . 158

20 Defining evidence in *t*-statistics 159
20.1 Example . . . 159
20.2 Evidence in the Student *t*-statistic . . . 159
20.3 The Key Inferential Function for Student's model . . . 163
20.4 Corrected evidence . . . 165
20.4.1 Matching *p*-values . . . 166
20.4.2 Reducing bias . . . 167
20.5 A confidence interval for the standardized effect . . . 169
20.5.1 Simulation study of coverage probabilities . . . 170
20.6 Comparing evidence in *t*- and *z*-tests . . . 170
20.6.1 On substituting s for σ in large samples . . . 170
20.7 Summary . . . 172

21 Two-sample comparisons 175
21.1 Drop in systolic blood pressure . . . 175
21.2 Defining the standardized effect . . . 176
21.3 Evidence in the Welch statistic . . . 177
21.3.1 The Welch statistic . . . 177
21.3.2 Variance stabilizing the Welch *t*-statistic . . . 178
21.3.3 Choosing the sample size to obtain evidence . . . 179
21.4 Confidence intervals for δ . . . 179
21.4.1 Converting the evidence to confidence intervals . . . 179
21.4.2 Simulation studies . . . 180
21.4.3 Drop in systolic blood pressure (continued) . . . 181
21.5 Summary . . . 181

22 Evidence in the chi-squared statistic 183
22.1 The noncentral chi-squared distribution . . . 183
22.2 A *vst* for the noncentral chi-squared statistic . . . 184
22.2.1 Deriving the *vst* . . . 184
22.2.2 The Key Inferential Function . . . 185
22.3 Simulation studies . . . 186
22.3.1 Bias in the evidence function . . . 186
22.3.2 Upper confidence bounds; confidence intervals . . . 187
22.4 Choosing the sample size . . . 190
22.4.1 Sample sizes for obtaining an expected evidence . . . 190
22.4.2 Sample size required to obtain a desired power . . . 192
22.5 Evidence for $\lambda > \lambda_0$ . . . 192
22.6 Summary . . . 193

23 Evidence in *F*-tests 195
23.1 Variance stabilizing transformations for the noncentral F . . . 195
23.2 The evidence distribution . . . 199
23.3 The Key Inferential Function . . . 202
23.3.1 Refinements . . . 205
23.4 The random effects model . . . 205

23.4.1 Expected evidence in the balanced case 207
23.4.2 Comparing evidence in REM and FEM 208
23.5 Summary 208

24 Evidence in Cochran's *Q* for heterogeneity of effects 209
24.1 Cochran's Q: the fixed effects model 210
24.1.1 Background material 210
24.1.2 Evidence for heterogeneity of fixed effects 212
24.1.3 Evidence for heterogeneity of transformed effects 213
24.2 Simulation studies 213
24.3 Cochran's Q: the random effects model 216
24.4 Summary 220

25 Combining evidence from *K* studies 221
25.1 Background and preliminary steps 221
25.2 Fixed standardized effects 222
25.2.1 Fixed, and equal, standardized effects 222
25.2.2 Fixed, but unequal, standardized effects 223
25.2.3 Nuisance parameters 223
25.3 Random transformed effects 224
25.3.1 The random transformed effects model 224
25.3.2 Evidence for a positive effect 225
25.3.3 Confidence intervals for κ and δ: K small 226
25.3.4 Confidence intervals for κ and δ: K large 226
25.3.5 Simulation studies 227
25.4 Example: drop in systolic blood pressure 229
25.4.1 Inference for the fixed effects model 231
25.4.2 Inference for the random effects model 232
25.5 Summary 232

26 Correcting for publication bias 233
26.1 Publication bias 233
26.2 The truncated normal distribution 235
26.3 Bias correction based on censoring 237
26.4 Summary 240

27 Large-sample properties of variance stabilizing transformations 241
27.1 Existence of the variance stabilizing transformation 241
27.2 Tests and effect sizes 242
27.3 Power and efficiency 245
27.4 Summary 249

Acknowledgements 251

Bibliography 253

Index 257

Preface

This work is called a *guide* because it is primarily a source of basic methods for scientists wanting to combine evidence from different experiments. It also promotes a deeper understanding of the notion of statistical evidence. Many scientists like to use p-values for this purpose, but evidence is obscured by the p-value. It is rather a measure of surprise: the smaller the p-value under the null hypothesis, the more untenable the null hypothesis becomes. As a simple measure for filtering out unremarkable experimental results, the p-value works remarkably well. But it is hard to interpret and combine across experiments, especially when one believes the null is false.

When one has in hand several 'significant' p-values from different experiments, all testing for the same effect, the conviction grows that an alternative hypothesis could be true. By considering the p-value as a random variable under alternative hypotheses, one sees that its distribution is highly skewed, making interpretation and comparison of p-values under alternative hypotheses unwieldy at best. By transforming the p-value with the probability integral transform $T(p) = \Phi^{-1}$ $(1 - p)$, where Φ is the standard normal cumulative distribution function, one obtains the *probit scale* under the null and a location change of it under alternative, centered on the expected evidence.

The consequences for interpretation of evidence are profound if one is in the habit of thinking of p-values as measuring evidence. A 'highly' significant p-value of 0.01 represents, on average, only about 40 % more evidence for the alternative than a significant 0.05 p-value, because $\Phi^{-1}(1 - 0.01) = 2.326$ and $\Phi^{-1}(1 - 0.05) = 1.645$. There is no conflict here. The p-values are computed under the null and are measures of surprise, while the evidence lies on a location probit scale. It will be seen that under alternatives, the evidence measures always have a normal distribution with standard deviation one. Thus evidence as defined here is a random quantity with a well-known distribution, and it has a standard error of one unit when estimating its expected value.

The above statements are strictly true only for the prototypical normal model with known standard deviation, but as demonstrated in the chapters to follow, many test statistics can be transformed onto the probit scale by means of variance stabilizing transformations. Each application requires its own special transformation, and the mathematical level required for applying them is minimal.

So what can the reader expect from this book? In Chapter 1 the main ideas on statistical evidence are introduced, to offer a taste of, and hopefully whet the appetite for, the methods and theory to come.

Part I illustrates how to interpret and combine statistical evidence for the simplest statistical problems. These methods come first, so those readers wanting quick access to the 'how to do it' can readily find what they want. The why and wherefore – the philosophy and theory – behind these guidelines are found in Part II, for those readers piqued by curiosity or skepticism.

Chapters 2–5 present methods for continuous measurements for which the normal model is deemed appropriate. Chapters 6–9 describe methods for discrete measurements for which binomial or Poisson models are adopted.

These two groupings are followed in Chapters 10 and 11 by two applications of chi-squared statistics, testing for goodness-of-fit and homogeneity. However, the emphasis is on finding evidence for the alternative hypotheses; for example, evidence *for* heterogeneity rather than evidence *against* homogeneity. Measuring heterogeneity is important in Chapters 12 and 13 wherein methods for combining evidence for effects from similar studies are presented. Chapter 13 gives methods for regression of evidence on covariates, and finally Chapter 14 shows how to account for publication bias.

All chapters in Part I have the same format: data, model, questions of interest, test statistic, transformation to evidence, interpretation, choosing sample size and confidence intervals. This general methodology in each chapter is followed by worked examples. Macros for the software package *R* which enable the reader to obtain these and results for other data are provided on the website http://www.wiley.com/go/meta_analysis.

Part II provides the motivation, theory and results of simulation experiments to justify the methodology. It is intended to be a coherent introduction to the statistical concepts required to understand our thesis that evidence in a test statistic can be calibrated when transformed to a canonical scale. This leads to an appreciation of the error inherent in evidence, and provides the foundation for combining evidence from different studies.

The spirit of this theory is akin to the Fisherian tradition in that it attempts to provide a basis for thinking about test statistics, but it differs from Fisher's significance testing in that evidence is calibrated under alternatives, not the null hypothesis. Links to the Neyman–Pearson tradition can be made, because the expected evidence is a sum of probits of false positive and false negative rates, from which an expression for the power function is realized. A totally different approach to evidence based on the likelihood function is provided by Royall (1997).

The chapters in Part II could easily be the basis for a statistics course for senior undergraduates, while students working through the examples in Part I will gain some experience with real data. It is recommended that all readers carefully study the first two chapters of both Parts I and II, before embarking on more adventurous selections.

Elena Kulinskaya
Stephan Morgenthaler
Robert G. Staudte

Part I

The Methods

1

What can the reader expect from this book?

Experiments are conducted. Data are gathered. Researchers are looking for an effect, a change predicted by their musings over a model. At the very least, they want to gauge the *direction* of change: how much evidence is there in the data for a positive effect? More, they want an estimate of the size of the effect.

The statistical evidence for the direction of change is found in a test statistic. But how does one define and measure this ‘statistical evidence’?

In this book we provide a theory for inference in which the word evidence is central and meaningful. We show how to transform test statistics from different studies onto the same calibration scale where it is easier to measure, interpret and combine the evidence in them. Our approach lays the foundation for a meta-analytic theory with *known* weights. Further, it often leads to accurate confidence intervals for standardized effects using smaller sample sizes than would be achieved through standard asymptotic approximations.

The coming chapters are divided into two parts, dealing with methods and theory, respectively. In this chapter we give a taste of things to come. After introducing the calibration scale for evidence, we apply the methods to data from the meta-analytic review literature. Then we discuss standardized effects, sometimes called effect sizes, for two-sample comparisons, and note that each standardized effect is a simple function of a correlation coefficient.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte
© 2008 John Wiley & Sons, Ltd

1.1 A calibration scale for evidence

1.1.1 *T*-values and *p*-values

Consider the simple normal model with unknown mean μ and standard deviation 1. Given n observations $X_1, \ldots, X_n$ one rejects the null $\mu = 0$ in favor of the alternative $\mu > 0$ if the sample mean $S = \bar{X}_n$ is 'large enough'. The test statistic S is known to contain the evidence required for the test, but the word evidence is rarely defined. In this case we define the *evidence for the alternative* to be the transformed statistic $T = \sqrt{n}\, S = \sqrt{n}\, \bar{X}_n$. This T is normally distributed with mean $\sqrt{n}\, \mu$ and standard deviation 1, so T is an unbiased estimator of its mean $\sqrt{n}\, \mu$ with standard error 1.

Note that the expected evidence $\sqrt{n}\, \mu$ grows linearly with μ, and we require that any definition of evidence for $\mu > 0$ would grow with μ. In addition, the expected evidence grows with the square root of the sample size; this is consistent with the notion from estimation that evidence for an unknown μ grows only at this rate: one needs four times as many observations to estimate μ with twice the accuracy, because the standard error of $\bar{X}_n$ is $1/\sqrt{n}$.

Thus evidence for the alternative as defined here is a random quantity which always has inherent error, in fact a standard normal error, whether or not the null hypothesis holds. If one observes $T = 1.645$ and reports this as evidence for the alternative, one should also note the standard error is 1; it is better to write 1.645 ± 1. When one does this, one realizes that what is sometimes called a 'significant' outcome could quite easily have been something else.

Now suppose that one has two independent experiments similar to the one above, with respective sample means $\bar{X}_1$ based on n_1 observations and $\bar{X}_2$ based on n_2 observations. How can we combine the evidence in $T_1 = \sqrt{n_1}\, \bar{X}_1$ and $T_2 = \sqrt{n_2}\, \bar{X}_2$ to obtain a single evidence T for the alternative $\mu > 0$? A good choice is $T_{\mathrm{comb}} = (\sqrt{n_1}\, T_1 + \sqrt{n_2}\, T_2)/\sqrt{n_1 + n_2}$, because it is the mean of all $n_1 + n_2$ observations, rescaled to have variance 1. Also, T is a linear combination of independent normal variables and hence normal, with expected evidence $\sqrt{n_1 + n_2}\, \mu$ and standard deviation 1. It is on the same calibration scale as T_1 and T_2. In particular if $n_1 = 9$, $n_2 = 16$ and $T_1 = 1.645$, $T_2 = 2.236$, then the combined evidence for $\mu > 0$ is $T_{\mathrm{comb}} = 2.848 \pm 1$.

Another way of combining the evidence for $\mu > 0$ is to take $(T_1 + T_2)/\sqrt{2}$ which is normal with mean $(\sqrt{n_1} + \sqrt{n_2})\, \mu/\sqrt{2}$ and variance 1. For the example in which $T_1 = 1.645$ and $T_2 = 2.236$ this combination yields 2.808 ± 1, which is slightly smaller than T_{comb}. Note that $\sqrt{n_1 + n_2}$ is always greater than or equal to $(\sqrt{n_1} + \sqrt{n_2})/\sqrt{2}$ and equality is achieved only when $n_1 = n_2$. Thus, the first combination of the evidence is on average always at least as good as the second one. The proof of the cited inequality is left to the reader; it follows from the concavity of the square root function.

Traditional 'significance' is only weak evidence for the alternative

So far we only have transformed the test statistic S onto a scale whose unit equals the standard deviation of $T = T(S)$. A traditional marker on this scale is 1.645, the point dividing ‘significant’ from ‘nonsignificant’ values. But of course there is almost no difference between the results $T = 1.644$ and $T = 1.646$, and adding and subtracting the true standard error of 1 puts evidence into its proper perspective: it has a standard normal error. The result $T = 1.645 \pm 1$ illustrates that $T = 1.645$ is unreliable. If forced to give an adjective describing such evidence, we would call it ‘weak’. Twice as much evidence, $T = 3.3$, will then be called ‘moderate’, and three times as much evidence, $T = 5$, will be called ‘strong’. See Figure 1.1 for plots of some evidence possibilities. These somewhat arbitrary descriptions are necessarily vague because evidence is a random quantity. But we think they are a more realistic guide than setting degrees of ‘significance’ based on p-values.

The p-value of an observed value of a test statistic is often thought to be a measure of evidence against a null hypothesis, with smaller values indicating larger evidence. In a certain sense this is true, but the p-value is conditional on the data from a particular experiment, and so has relevance only for that particular experiment. If one wants to compare p-values from different experiments, or even to combine the evidence in them as in meta analysis, one must take into account their distributional properties.

First assume the null hypothesis holds. Then the p-value, when considered as a random variable, is known to have a uniform distribution on the unit interval when the test statistic has a continuous distribution, and nearly uniform if the test statistic has a discrete distribution. So, one might argue, one can indeed combine p-values using

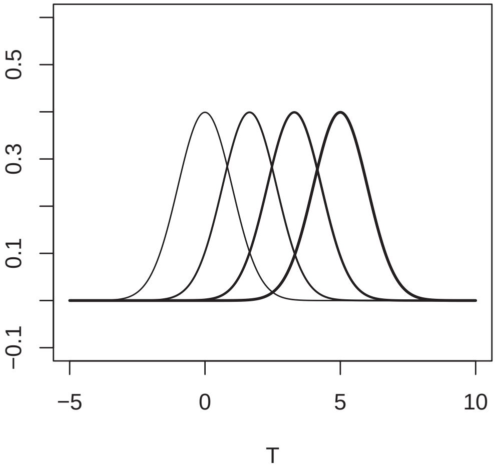


Figure 1.1 The distribution of evidence on the proposed calibration scale is always normally distributed with variance 1. When $\sqrt{n}\,\mu = 0$, the evidence T is centered on the origin; this is often called the null distribution of T. Other possibilities are centered on $\sqrt{n}\,\mu = 1.645$, 3.3 and 5; respectively shown from left to right. The point is that evidence is a random quantity with an unknown mean but standard normal error. Upon observing $T = 3.3$, one should report $T = 3.3 \pm 1$. This gives a clear indication not only of the magnitude, but also the error inherent in evidence T.

their common null uniform distribution, and assumed independence of experiments. But when one has in hand a number of small p-values, each of which is considered 'significant', the conviction grows that the null distribution is indeed false, and what is really desired is a combination of evidence that works whether or not the null hypothesis is true. Such a combination cannot be based on the assumption that the null hypothesis is true and that the p-value has a rectangular density. These considerations and others, explained in detail in Chapter 16, lead us to the conclusion that p-values, when considered as random variables, are on the wrong scale for calibration and interpretation of statistical evidence, and for forming a combined conclusion from a set of tests.

Before leaving this section we point out that a p-value for a test based on the T-statistic can be obtained if desired through the probit transformation of an observed value t of the evidence T. It is $p = 1 - \Phi(t) = \Phi(-t)$. For this simple example the p-values based on $T = T(S)$ are exactly the same as those based on S.

1.1.2 How generally applicable is the calibration scale?

So far we have only considered the simplest model of testing for a normal mean when the standard deviation is known. The transformation of the test statistic $S = \bar{X}_n$ to evidence $T = \sqrt{n}\, S$ only required multiplication by the square root of the sample size. In general one tries to select a transformation h of the test statistic S so that $T = h(S)$ is on this same unit normal calibration scale. In most routine problems of statistics this goal cannot be achieved completely, but it can be achieved approximately to a surprising degree for one- and two-sample binomial and Poisson models, for one- and two-sample t-tests and for chi-squared and F-tests. The first step then is to find the variance stabilizing transformation $h(S)$ for the particular model of interest, and the results of our and others' endeavors are presented in coming chapters.

In most cases the resulting evidence T is approximately normal with standard deviation 1 and mean which can be approximated $\mathrm{E}[T] \doteq \sqrt{n}\, \mathcal{K}(\delta)$. Here again n is the sample size, δ is a standardized effect and $\mathcal{K}$ is the Key Inferential Function. Knowing the Key is like knowing the power function in traditional Neyman–Pearson testing; it contains all the important information about the relationship between the standardized effect δ and its transformed value $\kappa = \mathcal{K}(\delta)$. This information can be exploited to choose sample sizes to obtain desired amounts of evidence, up to standard error 1, or to derive confidence intervals for δ.

Example 1. The one-sample *t*-test

Take $X_1, \ldots, X_n$ independent, each having the normal distribution with unknown mean μ and variance σ^2. The raw effect is $\mu - \mu_0$, where μ_0 is a known value determined by scientific context. The standardized effect is $\delta = (\mu - \mu_0)/\sigma$. For testing the null $\mu = \mu_0$ against the alternative $\mu > \mu_0$ the test statistic $S = \sqrt{n}\,(\bar{X}_n - \mu_0)/s_n$ is known to have a Student t-distribution with $n - 1$ degrees of freedom under the null hypothesis and a noncentral t distribution with the same $n - 1$ degrees of freedom and noncentrality parameter $\sqrt{n}\,\delta$. Chapter 20 contains further details, where it is also shown that a variance stabilizing transformation $T = h(S)$ has the property that, to a useful approximation, T has the $N(\sqrt{n}\,\mathcal{K}(\delta), 1)$ distribution for a wide range of values of n and δ.

The Key Inferential Function for this measure of evidence is

$$\begin{aligned}\mathcal{K}(\delta) &= \sqrt{2}\,\sinh^{-1}(\delta/\sqrt{2}\,)\\ &= \sqrt{2}\,\ln(\delta/\sqrt{2} + \sqrt{1+\delta^2/2}).\end{aligned}$$

This simple monotonic function together with the sample size n provide all the information required for inference regarding δ, provided n is not too small. For example, when $n = 10$ accurate 95 % confidence intervals can be derived for any δ satisfying $-2 < \delta < 2$.

Example 2. The one-sample binomial test

Let X have the binomial distribution with parameters n, p where n is known and $0 < p < 1$. For testing the null $p = p_0$ against the alternative $p > p_0$ it is customary to reject the null when the test statistic X is too large; or equivalently when $\hat{p} = X/n$ is too large. It is well known (see Chapter 18) that a classic transformation of $\hat{p}$ to the unit normal calibration scale is given by $T = h(\hat{p}) = 2\sqrt{n}\,\{\arcsin(\sqrt{\hat{p}}\,) - \arcsin(\sqrt{p_0}\,)\}$, and this transformation is improved if $\hat{p}$ is replaced by $\tilde{p} = (X + 3/8)/(n + 3/4)$. The Key Inferential Function for this transformation is

$$\mathcal{K}(p) = 2\big\{\arcsin(\sqrt{p}\,) - \arcsin(\sqrt{p_0}\,)\big\}.$$

This Key could have been expressed as a function of the raw effect $p - p_0$ or the standardized effect $\delta = \sqrt{n}\,(p - p_0)/\sqrt{p(1-p)}$ because these effects are monotonic functions of p, but for this example it would be an unnecessary notational complication. In Section 1.2 we illustrate how this arcsine transformation to the calibration scale can be employed to find and combine the evidence in several studies. But first we need to discuss several issues arising when considering more than one study on the same subject.

1.1.3 Combining evidence

Return to the simple normal model of Section 1.1.1, where we tacitly assumed that the true effect μ was the same for the two studies, instead of the more realistic assumption that $T_1 \sim N(\sqrt{n_1}\,\mu_1, 1)$, $T_2 \sim N(\sqrt{n_2}\,\mu_2, 1)$ where both μ_1, μ_2 are unknown. The joint null hypothesis is now $\mu_1 = 0 = \mu_2$, and there are many possible alternatives, each possibly requiring a different combination of evidence. For example, the alternative $\mu_w = (w_1\mu_1 + w_2\mu_2)/(w_1 + w_2) > 0$, for known positive weights w_1, w_2, suggests a combination $T_w = c(w_1T_1 + w_2T_2)$, with constant c chosen so that T_w has variance 1. And the joint alternative $\mu_1 > 0$ and $\mu_2 > 0$ suggests a combination of the form $T_{\text{min}} = h(\min\{T_1, T_2\})$ where h is a transformation to the unit normal calibration scale. The best combination for each alternative is a challenging problem in itself, which we do not pursue here. Rather we test or estimate an overall effect.

In traditional meta analysis it is common to assume the μ_k values are equal (the fixed effects model); or to assume that the μ_k values themselves are a random sample from a $N(\mu, \gamma^2)$ model (the random effects model), where γ^2 is a variance component introduced to explain the variability in μ_k. The advantage of these two models is that there is only one parameter of interest μ, the overall effect, and one can test hypotheses regarding μ or estimate μ without all the complications raised in the previous paragraph for fixed unequal effects.

More generally, we have K independent studies resulting in evidences T_k which are approximately normal with variance near 1 and $\mathrm{E}[T_k] \doteq \sqrt{n_k}\,\mathcal{K}(\delta_k)$ for $k = 1, \ldots, K$. Here T_k is the evidence for $\delta_k > 0$ based on n_k observations in the kth study, obtained by a suitable variance stabilizing transformation, and $\mathcal{K}$ is the associated monotonically increasing Key Inferential Function. There is a one-to-one correspondence between each δ_k and $\kappa_k = \mathcal{K}(\delta_k)$. The *fixed standardized effects model* in which all $\delta_k = \delta$ is easiest to deal with, because there is only one δ, hence one $\kappa = \mathcal{K}(\delta)$. One can find the evidence $T_{\text{comb}} = \sum_k \sqrt{n_k}\, T_k/\sqrt{N}$, where $N = \sum_k n_k$, as evidence for the alternative $\kappa > 0$, and hence also for $\delta > 0$. Note that $T_{\text{comb}} \sim N(\sqrt{N}\,\mathcal{K}(\delta), 1)$. One can also use $T_{\text{comb}} \pm z_{1-\alpha/2}$ to obtain a $100(1-\alpha)$ % confidence interval $[L, U]$ for κ, and by back-transformation for δ, namely $[\mathcal{K}^{-1}(L/\sqrt{N}\,), \mathcal{K}^{-1}(U/\sqrt{N}\,)]$.

In many problems the assumption that all $\delta_k = \delta$ is untenable, and testable using Cochran's Q test of homogeneity. In Chapter 24 a variant of Cochran's Q called Q^* is applied to the $\hat{\kappa}_k$'s to find the evidence T_{Q^*} for heterogeneity of the κ_k's and hence the δ_k's. On the basis of this evidence, the researcher may well prefer the following model.

The *random transformed (standardized) effects model* assumes that the κ_k's are a random sample of size K from the normal model with mean κ and variance γ^2, with both parameters unknown. Then the conditional distribution of each $\hat{\kappa}_k$, given κ_k, is $N(\kappa_k, 1/n_k)$, and unconditionally it is $N(\kappa, \gamma^2 + 1/n_k)$. Now when the n_k's are all equal, or when their reciprocals are negligible compared to γ^2, the $\hat{\kappa}_k$'s are just a sample of size K from a normal population with mean κ and common variance. Let $\bar{\kappa}$ and s_κ^2 denote the sample mean and variance of these transformed standardized effects. The usual t-test rejects the null $\kappa = 0$ in favor of $\kappa > 0$ when the statistic $S = \sqrt{K}\,(\bar{\kappa} - 0)/s_\kappa$ is large. The evidence in this statistic for $\kappa > 0$, and hence $\delta > 0$, is essentially $T = \sqrt{2K}\,\sinh^{-1}(S/\sqrt{2K}\,)$, as shown in Chapter 20.

If one desires to compute a confidence interval for δ, one can find a t-interval $[L, U]$ for κ first, namely $\bar{\kappa} \pm t_{K-1,1-\alpha/2}\, s_\kappa/\sqrt{K}$, and then $[\mathcal{K}^{-1}(L), \mathcal{K}^{-1}(U)]$ for δ by back-transformation.

1.2 The efficacy of glass ionomer versus resin sealants for prevention of caries

1.2.1 The data

The review by Ahovuo-Saloranta *et al.* (2004) contains three studies in which matching molar teeth in the same children formed the basis for paired comparisons. Two

Table 1.1 Summary of three studies by the authors shown. Note the evidence is in conflict, but this should not preclude an analysis; further studies may demonstrate that one sealant is superior to another. References to these three studies and more background can be found in Ahovuo-Saloranta *et al.* (2004).

		Resin sealant					
		+	−	+	−	+	−
Glass Ionomer	+	378	28	156	6	191	2
Sealant	−	3	3	37	7	9	1
		Arrow (1995)		Poulsen (2001a)		Poulsen (2001b)	

types of sealants were applied at random to the pair, and then the teeth were assessed after 24- to 44-month intervals to detect the presence '−' of one or more caries or '+' no caries. The results of these three studies are summarized in Table 1.1.

The *discordant pairs* are those for which the treatment and control responses differ; let f be the number of $(+, -)$ pairs and g be the number of $(-, +)$ pairs. In the first study there are $f = 28$ pairs for which the response was $(+, -)$: there were no caries in one tooth after glass ionomer treatment, while the corresponding tooth receiving resin sealant did have caries. There were $g = 3$ pairs in which the two treatments led to the opposite results $(-, +)$. The conditional distribution of f, given $f + g$ is binomial with parameters $f + g$ and p, where p is the probability that a discordant pair is $(+, -)$. A test of symmetry in treatment control outcomes is a test of $p = 0.5$, with alternative $p > 0.5$ corresponding to the treatment (in this case glass ionomer) having greater probability of '+' within a discordant pair. (See Lachin (2000), p. 180, for more details.) We can now compute the evidence for $p > 0.5$ in each of the three studies using $T = 2\sqrt{n}\,\{\arcsin(\sqrt{\tilde{p}}\,) - \arcsin(\sqrt{0.5}\,)\}$, where $\tilde{p} = (X + 3/8)/(n + 3/4)$.

1.2.2 Analysis for individual studies

1.2.2.1 Evidence for $p_k > 0.5$ in individual studies

In the first experiment, there are 31 discordant pairs, so conditionally, X_1 has the binomial$(31, p_1)$ distribution, where p_1 is the probability that glass ionomer is more effective than the resin sealant in preventing caries in the first experiment. The evidence against $p_1 = 0.5$ in favor of $p_1 > 0.5$ is $T_1 = 5.05$, displayed in column 3 of Table 1.2; this is what we would call 'strong' evidence.

In the second study, the distribution of X_2 is binomial$(43, p_2)$, where again, p_2 is the probability that glass iomomer is more effective in this study. The evidence against $p_2 = 0.5$ in favor of $p_2 > 0.5$ is $T_2 = -5.16$; that is, the evidence is even stronger than in the first study, but this time in the opposing direction.

Table 1.2 Summary of synthesis of evidence for the sealant data in Table 1.1.

k	X_k	n_k	$\tilde{p}_k$	T_k	$T_{1:k}$	L_k	U_k	$\hat{\kappa}_k$
1	28	31	0.903	5.05	5.05	0.763	0.976	0.548
2	6	43	0.140	−5.16	−0.67	0.057	0.265	−0.560
3	2	11	0.182	−2.12	−1.39	0.029	0.476	−0.230

For the third study, the evidence against $p_3 = 0.5$ in favor of $p_3 > 0.5$ is $T_3 = -2.12$, which is weak evidence for the alternative $p_3 < 0.5$. It is important to remember that all these evidence values have standard error 1.

Confidence intervals for p_k in individual studies

Confidence intervals $[L_k, U_k]$ for p_k are based on Equation (18.2), and for confidence 95% are shown in columns 7 and 8 of Table 1.2. Note that they are not centered on $\tilde{p}_k$, but are more reliable than intervals based on the standard asymptotic theory of adding and subtracting 1.96 standard errors to $\hat{p}$. For more details, see Chapter 18. These intervals suggest that the p_k are not equal, but nevertheless for completeness we assume this to be the case in the next section.

1.2.3 Combining the evidence: fixed effects model

If we were to assume that all $p_k = p$, then we could readily combine the evidence in the individual studies for $p > 0.5$. The results in column 6 of Table 1.2 are obtained sequentially: the entry in row k is based on the first k studies. The first two studies have strong conflicting evidence, and this is reflected by the combined evidence $T_{12} = (\sqrt{31}\,T_1 + \sqrt{43}\,T_2)/\sqrt{74} = -0.67$, shown in column 6. It is almost negligible. For the three studies, the combined evidence is $T_{1:3} = (\sqrt{74}\,T_{12} + \sqrt{11}\,T_3)/\sqrt{85} = -1.39$, which is quite weak evidence in favor of the resin sealant. Thus combining evidence on the calibration scale allows for cancelation of conflicting evidence, leading to the correct conclusion that there is no evidence for a common $p > 0$.

One can also obtain a confidence interval for p based on all three studies. Starting with the combined evidence $T_{1:3} = -1.39$, a 95 % confidence interval for the expected evidence $\sqrt{85}\,\mathcal{K}(p)$ is -1.39 ± 1.96, or $[L, U] = [-3.35, 0.57]$. Here the key is $\mathcal{K}(p) = 2\big\{\arcsin(\sqrt{p}\,) - \arcsin(\sqrt{0.5}\,)\big\}$, so $\mathcal{K}^{-1}(y) = \sin^2(y/2 + \pi/4)$. This leads to the 95 % interval $[\mathcal{K}^{-1}(L/\sqrt{85}\,)\,,\ \mathcal{K}^{-1}(U/\sqrt{85}\,)\,] = [0.32, 0.53]$ for p.

1.2.4 Combining the evidence: random effects model

The transformed effects $\hat{\kappa}_k = \mathcal{K}(\tilde{p}_k)$ are shown in Table 1.2, and their respective approximate normal $N(\kappa_k, 1/n_k)$ distributions depicted in Figure 1.2. The sample mean and standard deviation are $\bar{\kappa} = -0.081$ and $s_\kappa = 0.569$. A test for heterogeneity of these transformed effects based on Cochran's Q is described in Chapter 24, and

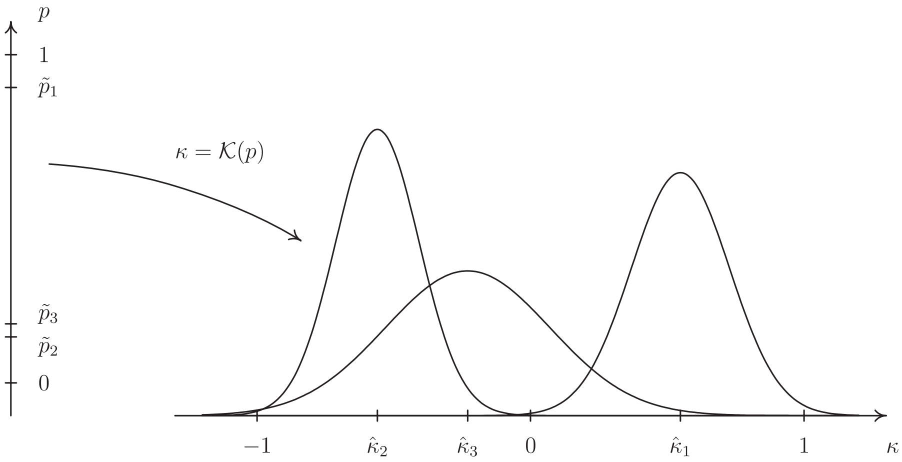


Figure 1.2 Transformation of the estimated probabilities that glass ionomer outperforms resin sealant into transformed effects $\hat{\kappa}_k = \mathcal{K}(\tilde{p}_k)$. The evidence T_k for a positive effect $p_k - 0.5 > 0$ has distribution that is approximately $N(\sqrt{n_k}\,\kappa_k, 1)$, so $\hat{\kappa}_k = T_k/\sqrt{n_k}$ has distribution that is approximately $N(\kappa_k, 1/n_k)$. These normal distributions are centered at respective unknowns κ_k's, and depicted in the plot centered at the respective estimates $\hat{\kappa}_k$'s.

the evidence for heterogeneity is strong ($T_{Q^*} \approx 4.5$) so a random transformed effects model is in order; it essentially adds a variance component to the model to account for the variability from study to study. Details are given in Section 25.3, where it is shown that if the reciprocals of the sample sizes are small compared to this component, then, even for a small number of studies K, the evidence for the overall $\kappa > 0$, and hence $p > 0.5$, is $T = \sqrt{2K}\,\sinh^{-1}(S/\sqrt{2K})$, where $S = \sqrt{K}\,(\bar{\kappa} - 0)/s_\kappa$.

For our data $S = -\sqrt{3}\,(0.081/0.569) = -0.25$ and so the evidence T for $\kappa > 0$, and hence $p > 0$, is negligible. (Note that here $T \approx S = -0.25$, because the function $\sinh^{-1}$ behaves like the identity near the origin.)

A confidence interval for a representative p can also be found, starting with the t-interval for κ of $\bar{\kappa} \pm t_{2,0.975} s_\kappa/\sqrt{3}$ or $[L, U] = [-1.49, 1.33]$. By transforming this interval back via $\mathcal{K}^{-1}(y) = \sin^2(y/2 + \pi/4)$, the 95 % confidence interval for p is $[0.002, 0.986]$. This interval tells us virtually nothing about p, but of course this is because the number of studies is small, and the results are contradictory. It confirms that the very strong assumption of a fixed effects model which led to the interval $[0.32, 0.53]$ for p is unwarranted.

1.3 Measures of effect size for two populations

For us an *effect size* is another term for *standardized effect*; that is, an effect divided by a suitable measure of scale. For a single population with mean μ, standard deviation σ, it is often taken to be the raw effect $\mu - \mu_0$ divided by σ. Here μ_0 is a hypothesized value of μ suggested by scientific context. The advantage of standardized effects over raw effects is that they are free of the units of measurement. For two populations with different variances σ_1^2, σ_2^2, the question arises of how to standardize the difference of their means $\Delta = \mu_1 - \mu_2$. The purpose of this section is to define a standardized effect δ for comparing two populations and its associated correlation effect size $\rho = \rho(\delta)$.

Let $X_1, \ldots, X_{n_1}$ be a sample of size n_1 from the first population and estimate μ_1 by the sample mean $\bar{X}$; similarly let $\bar{Y}$ be based on an independent sample $Y_1, \ldots, Y_{n_2}$ from the second population. Then an unbiased estimator of the effect $\Delta = \mu_1 - \mu_2$ is $\hat{\Delta} = \bar{X} - \bar{Y}$. Now, because $\hat{\Delta}$ is unbiased for Δ, the standard error $\mathrm{SE}[\hat{\Delta}]$ of $\hat{\Delta}$ satisfies

$$\left\{\mathrm{SE}[\hat{\Delta}]\right\}^2 = \mathrm{Var}[\hat{\Delta}] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \; .$$

Definition 1.1 *Let $N = n_1 + n_2$, and define the* standardized effect *by*

$$\delta = \frac{\Delta}{\sqrt{N}\, SE[\hat{\Delta}]} \; .$$

This effect size δ is free of the units of measurement. Note that δ is also free of the sample sizes, but does depend on the relative sample sizes, as well as Δ and unknown population variances.

There are numerous other definitions of effect sizes in the meta-analytic literature, including those that are Pearson product moment correlations between the variable of interest and a classification variable; this group includes the point-biserial correlation coefficient, see Cohen (1988) and Rosnow and Rosenthal (1996) and references therein. These measures of effect size are often called *correlation effect sizes* and will be denoted generically here by ρ. Each is related to a corresponding standardized effect δ by:

$$\rho = \frac{\delta}{\{1 + \delta^2\}^{1/2}} \; . \tag{1.1}$$

A plot of ρ against δ is shown in Figure 1.3. Note that $\rho = \rho(\delta)$ is a strictly increasing function of δ with inverse function $\delta = \delta(\rho) = \rho/\sqrt{1 - \rho^2}$. In addition, ρ is an odd function of δ; that is $\rho(-\delta) = -\rho(\delta)$ for all δ.

Examples

The above Definition 1.1 of standardized effect is employed directly in comparing two normal populations in Chapter 21. Another special case, comparing two Bernoulli populations, is also of interest, and discussed in Chapter 19. Here we reexpress the above results in a simpler notation for this problem. Assume each $X_i = 1$ or 0, respectively, with probabilities $p_1, 1 - p_1$; that is, X_i has the Bernoulli(p_1) distribution, and $\mu_1 = \mathrm{E}[X_i] = p_1$ and $\sigma_1^2 = p_1(1 - p_1)$. Similarly let each $Y_i \sim$Bernoulli(p_2). Then $\hat{p}_1 = \bar{X}$, $\hat{p}_2 = \bar{Y}$.

In this context $\Delta = p_1 - p_2$ and $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$. Further, letting $q = n_2/N$, where $N = n_1 + n_2$, and following the notation of Brown and Li (2005), let $p = qp_1 +$

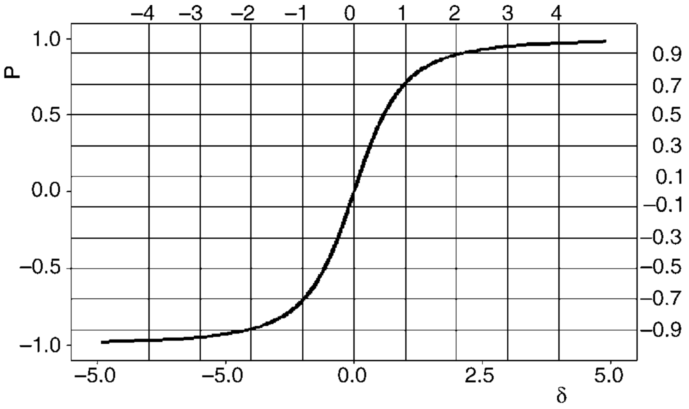


Figure 1.3 The graph of correlation effect size ρ against standardized effect δ.

$(1-q)p_2$. They observe that $N\text{Var}[\hat{\Delta}] = \zeta - \Delta^2$, where $\zeta = p(1-p)/\{q(1-q)\}$. The standardized effect is therefore $\delta = \Delta/\sqrt{\zeta - \Delta^2}$, and the associated correlation effect size is $\rho = \Delta/\sqrt{\zeta}$. The importance of this result to the theory presented here is that in Chapter 19 we define a new and effective variance stabilizing transformation for the risk difference $\Delta = p_1 - p_2$ and its associated Key Inferential Function is simply $\mathcal{K}(\rho) = \arcsin(\rho)$.

1.4 Summary

In this text we provide a unified theory of statistical inference in which the word ‘evidence’ is central and meaningful. It grows out of our conviction that the traditional ways of measuring evidence, in particular with probabilities, are neither intuitive nor useful when it comes to making comparisons between experimental results, or when combining them.

We measure evidence for an alternative hypothesis, not evidence against a null. To do this, we have in a sense adopted standardized scores for the calibration scale. Evidence for us is simply a transformation of a test statistic to another one (called *evidence*) whose distribution is close to normal with variance 1, and whose mean grows from 0 with the parameter as it moves away from the null. The transformation required depends on the model, and there is a rich legacy to draw upon from research in the last century.

The advantages of such a theory are many:

- **Conceptual simplicity.** Evidence T for an alternative is normally distributed with unknown mean and variance 1; it is an unbiased estimator of its mean that always has a standard normal error.

- **Usefulness.** The expected evidence often has the form $\mathrm{E}[T] \doteq \sqrt{n}\,\mathcal{K}(\delta)$, where $\mathcal{K}$ is a known Key Inferential Function. This formula facilitates finding sample sizes required to achieve desired amounts of evidence for an alternative, and deriving confidence intervals for δ.
- **Effectiveness.** Compared to methods based on standard asymptotics, these methods generally require smaller sample sizes to achieve good approximations (see Chapter 27).
- **Meta-analytic potential.** Combining evidence on this calibration scale is simpler, because it forms combinations of evidence with *known* weights.

Of course there are disadvantages, too, of which the reader is no doubt aware. One needs to become familiar with square root, arcsine and hyperbolic arcsine transformations. But in this opening chapter we have tried to convey the above listed potential benefits of defining evidence on the unit normal scale. We have sketched the ideas for the most important binomial and normal models, and illustrated the meta-analytic ideas on data from the recent review literature. We have concluded with some relations between effect sizes useful to us in comparing two populations.

2

Independent measurements with known precision

This chapter is a template for later chapters, and therefore should be read by all readers. It illustrates the methodology for the simplest normal model where only one parameter, the mean, is unknown, and the variance is known. In all subsequent sections a variance stabilizing transformation will be required to bring one onto this calibration scale.

2.1 Evidence for one-sided alternatives

Data and model

- We are given measurements $x_1, \ldots, x_n$ on a variable X obtained by an instrument of known precision.
- The measurements are regarded as independent observations which form a sample from a normal population with unknown mean μ and known standard deviation σ_0, the precision.

Question

- What is the evidence for an effect in a known direction? For example, what is the evidence against the null hypothesis $\mu = \mu_0$ and for the alternative $\mu > \mu_0$? Here the value μ_0 is known and determined by scientific context. The difference

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

$\mu - \mu_0$ is called the *effect*, while $\delta = (\mu - \mu_0)/\sigma_0$ is called the *standardized effect*.

- By symmetry, if alternatives $\mu < \mu_0$ are of interest, the problem is the same as for $\mu > \mu_0$, except for the direction. The change of direction is reflected in the sign of the evidence T. The former problem could be transformed into the latter by reflection about μ_0; i.e. replacing the deviations from the null value $x_i - \mu_0$ by its negative $\mu_0 - x_i$; or, replacing the observations x_i by $2\mu_0 - x_i$. Thus we only comment on the direction $\mu > \mu_0$ and interpret positive evidence as support for $\mu > \mu_0$.

Test statistic and distribution

- The usual test statistic is based on the arithmetic mean $\bar{x}_n = \sum_i x_i/n$ of the measurements; large values of $S = (\bar{x}_n - \mu_0)$ support the alternative $\mu > \mu_0$ over the null $\mu = \mu_0$. There is no natural boundary separating 'small' $\bar{x}_n$ from 'large', and that is why a calibration scale is desirable.
- The model for $\bar{x}_n$ is also normal, with the same unknown mean μ, but smaller variance σ_0^2/n. Also, the standard deviation of $\bar{x}_n$ is $\sigma_0/\sqrt{n}$ which is often called the *standard error* of $\bar{x}_n$.

Transformation to evidence

- Let the *evidence* be $T = \sqrt{n}\,(\bar{x}_n - \mu_0)/\sigma_0 = \sqrt{n}\,\hat{\delta}$. Then T will, on average, be equal to $\tau = \sqrt{n}\,\delta$. Also, the standard deviation of T is 1, and the values of T can be thought of as being drawn at random from a bell-shaped normal distribution. These facts can be summarized symbolically as $T \sim N(\sqrt{n}\,\delta, 1)$.

Interpretation

- The evidence T is an unbiased estimator of the expected evidence $\tau = E[T] = \sqrt{n}\,\delta$, with standard error $SE[T] = 1$. Therefore the evidence is closely related to τ, as shown in Figure 1.1. It displays the distribution of T for four values, $\tau = 0, 1.645, 3.3$ and 5, which in words we describe, respectively, as no evidence, weak, moderate and strong evidence for the alternative $\mu > \mu_0$. Note that there is a small amount of overlap in the use of these words.
- Under the null hypothesis $\mu = \mu_0$, the standardized effect $\delta = 0$, so $\tau = 0$ and T has the standard normal distribution with cumulative distribution function Φ. For an observed sample mean $\bar{x}_n$, the observed evidence is $T = \sqrt{n}(\bar{x}_n - \mu_0)/\sigma_0$ and the p-value is $p = 1 - \Phi(T) = \Phi(-T)$. Thus p-values can be recovered from T.
- The choice of $\tau = 1.645$ as the basic unit of calibration is for compatibility with the well-established $p = 0.05$ in significance testing; while this boundary traditionally separates 'significant' from nonsignificant results, all scientists

know this boundary is arbitrary and in terms of evidence it is *weak*. It is weak partly because when an experiment has just achieved a boundary result of 0.05, the expected p-value in an independent replication of the experiment is 0.12 (see Section 16.2.2). It is also weak because it is unreliable in that the standard error of T is 1, and relative to the size of $T = 1.645$, this standard error is large.

- The relative error in T, $\mathrm{SE}[T]/\mathrm{E}[T] = 1/\tau$, becomes smaller and smaller as τ increases. Because $\tau = \sqrt{n}\,\delta$, choosing the sample size to achieve a desired expected evidence for a relative effect of interest becomes an option. Another is combining the evidence from several experiments.

- For any fixed n and $\mu > \mu_0$ which determine the expected evidence $\tau = \sqrt{n}\,\delta$, one needs to increase n by a factor of 4 to move the density of T located at τ to 2τ and by a factor of 9 to move it from τ to 3τ. In particular, if the expected evidence is weak, $\tau = 1.645$, then 4 times as much work will yield moderate evidence of 3.3, and 9 times as much work is required for strong evidence of 5.

- The question arises as to what to do with negative values of T. They could be set equal to 0, because they are in a direction contrary to the alternative $\mu > 0$. However, we view evidence for such one-sided alternatives as the first step in finding evidence for two-sided alternatives, which are usually advocated. And preserving the direction of evidence through the sign means that when combining evidence in several studies, contradictory results are allowed to cancel each other out. Not to preserve the sign is to throw away valuable information. For further discussion of this question and the above remarks, see Section 2.2 and Chapter 16.

Choosing the sample size to achieve a desired amount of evidence

- If one wants the evidence for a particular standardized effect of scientific interest, call it $\delta_1 = (\mu_1 - \mu_0)/\sigma_0 > 0$, to be τ, one needs to solve $\tau = \sqrt{n}\,\delta_1$ for the sample size n. For example, to obtain ‘strong’ expected evidence $\tau = 5$ one requires $n = 25/\delta_1^2$. This does not *guarantee* strong evidence T, because T has standard error 1.

- Let z_α denote the α *quantile* of the standard normal model; it satisfies $\Phi(z_\alpha) = \alpha$. For those steeped in the Neyman–Pearson tradition, it is of interest to compare the above choice of n with that needed to obtain power $1 - \beta$ of detecting δ_1 at level α; it satisfies $\sqrt{n}\,\delta_1 = z_{1-\alpha} + z_{1-\beta}$. Hence the relationship between expected evidence τ, level α and power $1 - \beta$ is

$$\tau = z_{1-\alpha} + z_{1-\beta}. \tag{2.1}$$

What is usually asked for is power 0.8 at level $\alpha = 0.05$, and this corresponds to expected evidence $\tau = z_{0.95} + z_{0.8} = 1.645 + 0.842 \approx 2.5$, which is between weak and moderate. To obtain moderate evidence for alternative δ_1 at level 0.05 one needs power 0.95 of detecting it, not 0.8. To obtain strong evidence of $\tau = 5$ and maintain $\alpha \le \beta(\delta_1)$, one can take $\alpha = \beta = 0.005$, say.

Confidence intervals

- Let $c = z_{1-\alpha/2}$. Then a $100(1-\alpha)\,\%$ confidence interval for τ is $[T-c, T+c]$. It follows that with the same confidence $\delta = \tau/\sqrt{n}$ lies in the interval $[(T-c)/\sqrt{n}\,, (T+c)/\sqrt{n}\,]$; and for the effect $\mu - \mu_0$ the interval is

$$\left[\frac{\sigma_0(T-c)}{\sqrt{n}}\,,\ \frac{\sigma_0(T+c)}{\sqrt{n}}\right]. \tag{2.2}$$

Usually 95 % confidence is desired, and for this case $c = z_{0.975} = 1.96$.

2.2 Evidence for two-sided alternatives

In many, if not most, applications in which the measurements are modeled by a symmetric distribution, the researcher does not have enough prior knowledge to make the very strong assumption that the alternative to the null $\mu = \mu_0$ can only be in a specific direction. And doing so in the case of testing for the mean of a symmetric distribution means the *p*-value is only one-half what it would be if the two-sided alternative $\mu \neq \mu_0$ were specified; thus the evidence against the null is overstated. Such action is especially notable if an 'insignificant' 0.1 result is presented as a 'significant' 0.05, and hence strenuous objection to assuming one-sided alternatives is frequently made.

While we agree with this objection, it is equally important to keep in mind that when combining evidence from different studies, the *direction* as well as the magnitude of evidence needs to be known, so that conflicting findings are not hidden and can be accounted for. We therefore recommend reporting both one-sided and two-sided evidence.

Data, model and test statistic

Exactly as in Section 2.1.

Question

- What is the evidence for $\mu \neq \mu_0$?

Conversion to evidence

- Let $c = z_{0.75}/\sqrt{2} = 0.6745/\sqrt{2} = 0.477$. The evidence for the two-sided alternative $\mu \neq \mu_0$ to the null $\mu = \mu_0$ is motivated in Section 17.4.1 and defined in terms of the absolute value of evidence $|T|$ for one-sided alternatives by

$$T^{\pm} = \begin{cases} \sqrt{T^2 - c^2} - c, & \text{for } |T| \geq z_{0.75}; \\ c - \sqrt{\left[\Phi^{-1}(1.5 - \Phi(|T|))\right]^2 - c^2}\,, & \text{for } |T| < z_{0.75}. \end{cases} \tag{2.3}$$

Interpretation

- Evidence of 1.645 or -1.645 for a one-sided alternative, corresponding to a one-sided p-value of 0.05, is converted into two-sided evidence of 1.10, corresponding to the two-sided p-value of 0.136.
- The difference $|T| - T^{\pm}$ is the amount of evidence one loses for assuming a two-sided alternative when there is prior knowledge to assume a direction; it is also the amount that the evidence is overstated, by assuming a one-sided alternative when there is no justification for doing so. While this amount is not negligible, it is much smaller than 'halving the p-value' would suggest. For values of $|T|$ bigger than about 1.5, this turns out to be approximately the constant value of $c = z_{0.75}/\sqrt{2} = 0.477$.
- One-sided evidence can be positive or negative, indicating support for $\mu > 0$ or $\mu < 0$, respectively. Since we always want evidence to be roughly normally distributed, the same must hold for evidence for a two-sided alternative, even though negative values for the evidence can no longer be interpreted as giving evidence in the opposite direction. A negative value of the evidence for two-sided alternatives simply indicates that none of the alternatives is more convincing than the null value.

2.3 Examples

Measurements with known precision are common in manufacturing, where the consumer wants to know if a product meets the standard claimed on the label. A regulatory agency can take a random sample of a product under investigation, and look for evidence that the product complies with the rules. A manufacturer meanwhile will institute quality control procedures to ensure compliance.

When storing blood samples, do the concentrations of key markers change over time or do they remain stable? This can be checked with an experiment where two measurements are taken, one using fresh samples and the other after a period of storage. Whether these two results are close to each other then becomes the question of interest.

Determining whether a person is driving under the influence of an illegally high blood alcohol content is yet another example. Measurements always vary, and if the precision is known, the sample mean summarizes the available evidence. How does one calibrate this evidence?

2.3.1 Filling containers

A paint manufacturer fills 10 liters of white paint into cans that hold as much as 10.5 liters. The amount of paint the filling machine squirts into each can varies, and this inherent variability has known standard deviation $\sigma_0 = 0.2$ liters. The actual amount of paint in a sealed can is determined by net weight and conversion of weight to volume; these measurements are highly accurate and can be taken as exact for our purposes. Suppose the manufacturer is subject to regulatory fines if a random sample of four cans is found to have a mean volume of less than 9.67 liters. How was this value determined and how would we judge it from the point of view of evidence?

If a random sample of four cans leads to an average exactly equal to the regulatory limit $\bar{x}_4 = 9.67$, the p-value turns out to be $p = 0.0005$. Thus it can be argued that if the manufacturer is actually complying with the regulations, there is only 1 chance in 2000 of mistakenly charging fines for noncompliance. There is nothing wrong with this calculation. But if one were to add that $p = 0.0005$ is 'very strong evidence' in favor of the average filling volume being less than 10 liters, we would object to the words in quotes. The p-value of 0.0005 sounds impressively small, and is only 1/100 of the 'significant' 0.05. But is it 100 times more evidence against the null?

The evidence is $T = \sqrt{n}(\bar{x}_n - \mu_0)/\sigma_0 = 2 \times (9.67 - 10)/0.2 = 3.3$. We suggest that the statistic T is a better measure of evidence for the alternative, and -3.3, corresponding to 0.0005, is only twice the size of -1.645, corresponding to 0.05. On the probit calibration scale, the outcome $T = -3.3$ is seen to be moderate evidence for the alternative $\mu < 10$ rather than very strong evidence.

2.3.2 Stability of blood samples

This example is from Brown and Hollander (1977). The variable of interest is the level of triglyceride in blood plasma. Two measurements are taken, one on a fresh sample and the second one after 8 months in frozen storage. The concentration is expressed in mg/100 ml and it is known that the standard error of the analytic technique is equal to 4. For the difference x of two independent measurements, this results in a standard error of $\sqrt{4^2 + 4^2} = \sigma_0 = 5.7$. The rounded differences x_i before and after storage of $n = 30$ blood samples are:

−8	5	−4	−4	−1	1	8	8	−9	6	2	−2	7	−3	1
7	−3	−4	−2	−5	−2	5	6	3	−4	−1	14	−2	1	13

From this we find $\bar{x}_{30} = 1.1$. The corresponding one-sided evidence against $\mu = 0$ is $T = \sqrt{3}0(1.1 - 0)/5.7 = 1.1$, which is not even large enough to earn the qualifier of 'weak evidence'.

The 95% confidence interval for μ is

$$\left[\frac{\sigma_0(T - 0.477)}{\sqrt{n}}, \frac{\sigma_0(T + 0.477)}{\sqrt{n}}\right] = [-0.94, 3.14], \tag{2.4}$$

and thus quite wide.

2.3.3 Blood alcohol testing

Blood alcohol testing of drivers involved in accidents or even selected 'at random' is a legal requirement in many countries. An in-depth review of several methods by

Devleeschouwer *et al.* (2004) provides estimates of the precision of these methods. The variable of interest is X, the blood alcohol content in grams/liter.

As is often the case when measuring positive amounts, observations on blood alcohol have approximately constant coefficient of variation $\gamma = \sigma/\mu$, where σ is the standard deviation of the measurements and μ is the true blood alcohol content. For this same reason analytical chemists prefer to express the precision of their observations in terms of the coefficient of variation.

Because $\sigma = \gamma\mu$ is a percentage of μ, we cannot apply the test discussed in this chapter immediately. The link between μ and σ suggests the use of the logarithmic transformation. To see why, write $X = \mu \times (X/\mu)$ and note that X/μ has expected value of 1 and variance γ^2. Taking the logarithm leads to

$$\ln(X) = \ln(\mu) + \ln(X/\mu) = \ln(\mu) + \ln[1 + (X - \mu)/\mu] \approx \ln(\mu) + (X - \mu)/\mu,$$

where we have used the approximation $\ln(1 + u) \approx u$ for small u. The above equation also shows that $\ln(X)$ has expected value $\ln(\mu)$ and variance γ^2. The logarithmic

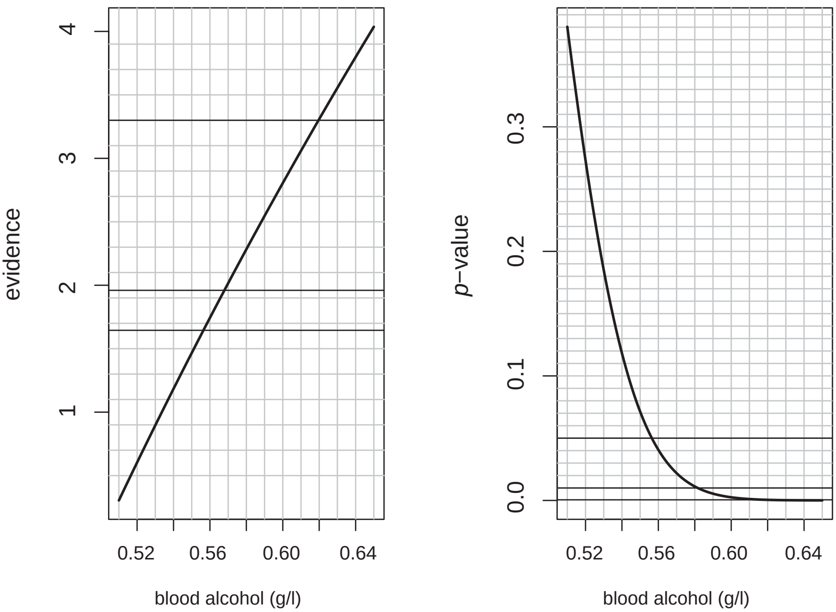


Figure 2.1 Plot of evidence and p-value against average of four readings on a breathalyzer test. The evidence for the alternative is weak at $T = 1.645$ when $\bar{x}_4 = 0.56$, and this corresponds to the p-value of 0.05. When $T = 2.32$, which is only slightly larger given that the standard error of T is 1, the p-value is 0.025, and when $T = 3.3$, twice the weak value, the p-value is 0.0005. Both plots are correct; but the interpretation is different, because the p-value plot assumes the null hypothesis $\mu = 0.5$ is true. The plot of T simply assumes μ is unknown, and comes with the proviso that T has standard error 1.

transformation of an observation with constant coefficient of variation will approximately have constant standard deviation. This is an example of a variance stabilizing transformation.

A subject involved in an accident must take four independent readings on a test and these lead to a sample mean of $\bar{x}_4$. This statistic is used to test the null hypothesis $\mu = 0.5$ grams/liter against the alternative $\mu > 0.5$, with the null rejected if $\bar{x}_4$ is large enough. The standard deviation of $\bar{x}_4$ is $\sigma/\sqrt{4}$, whereas its expected value remains equal to μ. The coefficient of variation of $\bar{x}_4$ is thus equal to $\gamma/\sqrt{4}$. Applying the above variance stabilizing transformation shows that $\ln(\bar{x}_4)$ has expected value $\ln(\mu)$ and standard deviation $\gamma/\sqrt{4}$.

It is known that $\gamma = 0.13$ for a certain blood testing kit. Thus $\ln(\bar{x}_4)$ has approximate standard error $0.13/\sqrt{4}$. Hence the p-value of an observed $\bar{x}_4$ is

$$p = 1 - \Phi\left(\sqrt{4}\{\ln(\bar{x}_4) - \ln(0.5)\}/0.13\right).$$

This p-value is plotted as a function of $\bar{x}_4$ in Figure 2.1, along with a plot of the evidence for the alternative $T = \sqrt{4}\{\ln(\bar{x}_4) - \ln(0.5)\}/0.13$.

Note that the evidence rises almost linearly with the sample mean (and would be exactly linear if we had not needed to use the log scale). But the p-value is hard to read and interpret in the region where it becomes small and is considered significant.

3

Independent measurements with unknown precision

For normal models with both parameters unknown, one may be interested in making inferences regarding μ, treating σ as a nuisance parameter, or σ with μ as the nuisance, and traditional methods based on the Student t-statistic or chi-squared statistic are available. The inference for μ is studied here along with inference for $\delta = (\mu - \mu_0)/\sigma$, the *standardized effect*. The evidence for the one-sided alternative $\mu > \mu_0$ is equivalent to $\delta > 0$, because $\sigma > 0$.

3.1 Effects and standardized effects

Data and model

- Given measurements $x_1, \ldots, x_n$ on a variable X obtained by an instrument of unknown precision.
- The measurements are considered independent observations from a normal population with unknown parameters μ, σ.

Questions

- What is the evidence for a positive effect $\mu - \mu_0 > 0$; or, equivalently, for a positive standardized effect $\delta > 0$?
- What is a confidence interval for μ or for δ?

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

Test statistic and distribution

- The Student t-statistic $t_{n-1} = \sqrt{n}\,\hat{\delta}$, where $\hat{\delta} = (\bar{x}_n - \mu_0)/s_n$ is an unbiased estimator of δ and $\bar{x}_n$ and $s_n^2 = \sum_i (x_i - \bar{x}_n)^2/(n-1)$ are the usual sample mean and variance. Larger values of the test statistic favor the alternative $\delta > 0$ over the null $\delta = 0$, and we want to transform the statistic onto the probit calibration scale.

- The distribution of t_{n-1} is the noncentral $t_\nu(\lambda)$ distribution with $\nu = n - 1$ degrees of freedom (hereafter abbreviated df) and noncentrality parameter $\lambda = \sqrt{n}\,\delta$. Under the null hypothesis, $\lambda = 0$ and t_{n-1} has the familiar central Student t-distribution, which approaches the standard normal with increasing n.

Transformation to evidence

- It turns out that t_{n-1} can be transformed to evidence $T = \sqrt{n}\,\mathcal{K}(\hat{\delta})$ having an approximate normal distribution with mean $\tau = \sqrt{n}\,\mathcal{K}(\delta)$ and variance 1, where $\mathcal{K}$ is given by Equation (3.1), for sample sizes $n \geq 5$ and δ encountered in applications.

- A modification T_{unbiased} of $\sqrt{n}\,\mathcal{K}(\hat{\delta})$ is analyzed in Section 20.4.2. It is $T_{\text{unbiased}} = \sqrt{n}\,\hat{\mathcal{K}}_{\text{unbiased}}$, with $\hat{\mathcal{K}}_{\text{unbiased}}$ defined by Equation (20.8). The corrected evidence T_{unbiased} is preferable to T. Its performance improves with sample size, as suggested by the following guidelines:

  - For $n = 5$ and $|\delta| < 2$ the variance is stabilized near 0.85, but nominal 95 % confidence intervals for δ have coverage nearer 97 %.
  - For $n = 10$ and $|\delta| < 10$ the variance is stabilized near 1.0, and nominal 95 % confidence intervals for δ are reliable for $|\delta| < 2$. This interval includes most δ encountered in applications.
  - For $n = 25$ and $|\delta| < 10$ the variance is stabilized near 1.0, and nominal 95 % confidence intervals are reliable for $|\delta| < 10$.

Interpretation

- The crucial ingredient $\mathcal{K}$ which determines the expected evidence is defined for each δ by

$$\mathcal{K}(\delta) = \sqrt{2}\,\ln\left(\frac{\delta}{\sqrt{2}} + \sqrt{1 + \frac{\delta^2}{2}}\right), \tag{3.1}$$

where $\ln(x) = \log_e(x)$ is the natural logarithm. The formula (3.1) for $\mathcal{K}(\delta)$ looks complicated, but it has a simple graph, as shown in Figure 3.1. Some values of $\tau = \sqrt{n}\,\mathcal{K}(\delta)$ are given in Table 3.1.

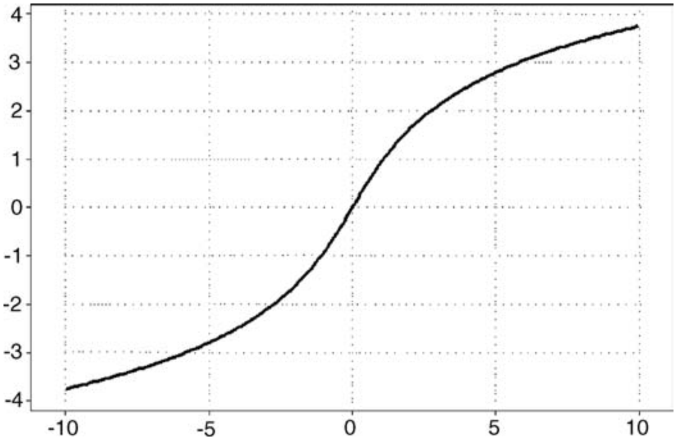


Figure 3.1 $\mathcal{K}(\delta)$ plotted as a function of δ. The graph is typical of many key functions $\mathcal{K}(\cdot)$ which determine the expected evidence in different contexts, in that for small values δ the function $\mathcal{K}(\delta) \approx \delta$, but larger values are diminished in magnitude, in this case logarithmically. Thus the expected evidence $\sqrt{n}\,\mathcal{K}(\delta)$ is not usually a linear function of δ, except for the model of Chapter 1.

- The approximate power function of the Student t-test can be obtained from T as follows, using the normal model $N(\tau, 1)$, where $\tau = \sqrt{n}\,\mathcal{K}(\delta)$, as an approximation to the distribution of T. A level-α test rejects the null $\delta = 0$ when $T \geq z_{1-\alpha}$. Let $\beta(\delta_1)$ denote the probability of falsely accepting the null when δ_1 is the true alternative. Then the power of the level-α test for detecting an alternative $\delta_1 > 0$ is

$$\begin{aligned} 1 - \beta(\delta_1) &= P_{\delta_1}(T \geq z_{1-\alpha}) \\ &= \Phi(\tau - z_{1-\alpha}) \\ &= \Phi(\sqrt{n}\,\mathcal{K}(\delta_1) - z_{1-\alpha}). \end{aligned} \tag{3.2}$$

Table 3.1 The second row contains some values of the monotonically increasing function $\mathcal{K}(\delta)$. The expected evidence in the Student t-statistic for the alternative $\delta > 0$ is $\tau = \sqrt{n}\,\mathcal{K}(\delta)$; examples for $n = 5$ and $n = 10$ are also tabled. Strong expected evidence of $\tau = 5$ for $\delta = 2$ is possible with sample size 10.

δ	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00
$\mathcal{K}(\delta)$	0.49	0.93	1.31	1.62	1.89	2.12	2.32	2.49
$\sqrt{5}\,\mathcal{K}(\delta)$	1.10	2.08	2.92	3.62	4.22	4.73	5.18	5.57
$\sqrt{10}\,\mathcal{K}(\delta)$	1.55	2.94	4.13	5.13	5.97	6.69	7.32	7.88

This formula can be rewritten to give the expected evidence in terms of level and power $\tau = z_{1-\alpha} + z_{1-\beta(\delta_1)}$. Here $\mathcal{K}(\delta)$ is given by (3.1), but these are general relationships between expected evidence, level and power for any Neyman–Pearson test based on a statistic which can be variance stabilized and normalized simultaneously. Usually it will be a rough approximation, but it works well for the t-test (see the discussion in Example 5 of Section 27.3).

Choosing the sample size to achieve a desired amount of evidence

- To obtain expected evidence $\tau = \sqrt{n}\,\mathcal{K}(\delta_1)$ for a standardized effect of scientific interest when this effect actually exists, one needs to solve $\tau = \sqrt{n}\,\mathcal{K}(\delta_1)$ for the sample size n; that is, $n = \{\tau/\mathcal{K}(\delta_1)\}^2$. This sample size calculation works well for $n \geq 10$ (see Table 27.2).

Confidence intervals

- Let $c_n = t_{n-1,0.975}$ be the 0.975 quantile of the Student t-distribution with $n - 1$ df. A 95 % confidence interval for μ is given by

$$[L, U] = \left[\bar{x}_n - c_n \frac{s_n}{\sqrt{n}}\ ,\ \bar{x}_n + c_n \frac{s_n}{\sqrt{n}}\right]. \qquad (3.3)$$

- Let $c = z_{0.975} = 1.96$. A nominal 95 % confidence interval for δ is given by

$$[L, U] = \left[\mathcal{K}^{-1}\left(\frac{T - c}{\sqrt{n}}\right), \mathcal{K}^{-1}\left(\frac{T + c}{\sqrt{n}}\right)\right], \qquad (3.4)$$

where $\mathcal{K}^{-1}(y) = \{e^{y/\sqrt{2}} - e^{-y/\sqrt{2}}\}/\sqrt{2}$ is the inverse function to $y = \mathcal{K}(\delta)$. The coverage of this interval is good when $n \geq 10$ and $|\delta| < 2$, the range usually encountered in applications. The range of good coverage improves with n; for example when $n = 25$ the range can be extended to $|\delta| < 10$.

3.2 Paired comparisons

Data and model

- Given pairs of measurements $(x_1, y_1), \ldots, (x_n, y_n)$ on a variable pair (X, Y), where the pairing is often deliberate to remove some other factor through differencing.
- The differences $d_i = y_i - x_i$ for $i = 1, \ldots, n$ are considered independent observations from a normal population with unknown parameters μ_d, σ_d^2. Here the X, Y variables are usually correlated. Each pair is a block within a randomized block design when the assignment of subjects within each pair is at random: one to receive treatment, the other serving as control. Then X, say, measures the control outcome and Y the treatment outcome.

Questions

- What is the evidence for a positive difference $\mu_d > 0$?
- Or, equivalently, for $\delta_d = \mu_d/\sigma_d$, what is the evidence for a positive standardized effect $\delta_d > 0$?

Test statistic and distribution

- The Student t-statistic $t_{n-1} = \sqrt{n}\,(\bar{d}_n - \mu_d)/s_n$, where $\bar{d}_n$ and $s_n^2 = \sum_i (d_i - \bar{d}_n)^2/(n-1)$ are the usual sample mean and variance of the d_i's. Larger values of the test statistic favor the alternative $\mu_d > 0$ over the null $\mu_d = 0$.
- The distribution of t_{n-1} is the noncentral $t_\nu(\lambda)$ distribution with $\nu = n - 1$ df and noncentrality parameter $\lambda = \sqrt{n}\,\delta$.

Conversion to evidence, interpretation and confidence intervals

- The evidence T_d in the t-statistic based on the differences is obtained as in Section 3.1. Confidence intervals for μ_d or δ_d are found using (3.3) and (3.4) of Section 3.1, with T_d replacing T.
- Only the interpretation changes, because two variables are involved. With differences defined by $d_i = y_i - x_i$, positive evidence T_d measures the support for the Y variable exceeding the X variable. The confidence interval for μ_d captures the size of the mean difference, while the confidence interval for δ_d captures the size of the mean difference *relative to the precision of the differences.*

3.3 Examples

These examples compare data summarized in a one-sample t-statistic with a fixed boundary of scientific interest. The second arises as a result of taking differences of paired observations, so the boundary or null hypothesis is 0.

3.3.1 Daily energy intake compared to a fixed level

The average daily energy intake in kilojoules (kJ) of 11 healthy women is compared to a standard recommended intake level of 7725 kJ in a study by Manocha *et al.* (1986) and also analyzed in Altman (1991). The 11 observations are, after ordering,

5260, 5470, 5640, 6180, 6390, 6515, 6805, 7515, 7515, 8230, 8770.

A normal model with unknown parameters μ, σ^2 is proposed for testing the null hypothesis $\mu = \mu_0 = 7725$ against $\mu < \mu_0$ or $\mu \neq \mu_0$. The sample mean and standard deviation are $\bar{x} = 6753.6$ and $s = 1142.1$. Thus the t-statistic is $t = \sqrt{11}\,(\bar{x} - \mu_0)/s = -2.821$, which supports the one-sided alternative $\mu < \mu_0$ with a p-value of 0.009 and the two-sided alternative with p-value 0.018. A two-sided t-interval for the effect $\mu - \mu_0$ is obtained by subtracting μ_0 from (3.3) and equals $[-1738.7, -204.1]$. These are the traditional ways of summarizing the data. But they do not give the evidence for the one- or two-sided alternatives, nor a confidence interval for $\delta = (\mu - \mu_0)/\sigma$, the mean effect, relative to the population standard deviation.

By transforming the t-statistic with (20.4) one obtains evidence for the one-sided alternative of $T = -1.947 \pm 1$ and for the two-sided alternative of $T^{\pm} = 1.63 \pm 1$. The standard errors are recorded to emphasize the error in measuring evidence. The evidence in this experiment for the two-sided alternative is weak, which is not unusual when the p-values are in the 0.01 to 0.05 range.

The relative mean effect δ is a measure of how the dietary intake differs from a recommended level in units σ which are particular to the population of interest, and is free of the units of measurement. This is arguably a more useful concept than the raw effect $\mu - \mu_0$, unless one has a good understanding of kilojoules. The confidence interval for δ obtained from (3.4) is $[-1.470, 0.004]$.

3.3.2 Darwin's data on *Zea mays*

Measurements on the plant *Zea mays* were collected by Charles Darwin and analyzed by Fisher (1935). As reported by Manly (1991), Darwin took 15 pairs of plants where within each pair the two plants 'were of exactly the same age, were subjected from the first to last to exactly the same conditions, were descended from the same parents'. One individual in each pair was cross-fertilized and the other was self-fertilized. The heights (x_i, y_i) for the pair of offspring were then measured to the nearest eighth of an inch over 12 inches. The original data are shown in the next section; here we just list the differences $d_i = x_i - y_i,\ i = 1, \ldots, 15$:

$$49, -67, 8, 16, 6, 23, 28, 41, 14, 29, 56, 24, 75, 60, -48.$$

The question of interest to Darwin was whether these results confirm the general belief that the offspring from crossed plants are superior to those from self-fertilized in the sense of having greater mean height. Thus we want to test $\mu_1 = \mu_2$ and the general belief is the one-sided alternative $\mu_1 > \mu_2$. However, it is possible that $\mu_1 < \mu_2$, so evidence for both one- and two-sided alternatives will be calculated.

The Student t-statistic $t_{14} = \sqrt{15}\,(\bar{d} - 0)/s_d = 2.148$ has 14 df so the one-sided p-value is found to be 0.025 and the two-sided p-value is therefore 0.05. A 95 % confidence interval for the mean difference is $[0.03, 41.84]$. Manly (1991) also explains how to compute p-values and confidence intervals for μ_d using permutation arguments which do not require the assumption of a normal model.

We next find the evidence in t_{14} for the one-sided alternative by transformation to evidence $T_d = 1.73$ for the one-sided alternative $\delta_d > 0$. By Equation (2.3) the evidence is $T_d^{\pm} = 1.38$ for the two-sided alternative $\delta_d \neq 0$. Each of these measures of evidence is best reported together with their standard errors 1.73 ± 1 and 1.38 ± 1. While some may lament the fact that these values are weak evidence with error in them, it is more realistic than reporting 0.025 and 0.05 as 'significant' measures of evidence with no error in them. These latter numbers are of course, correct p-values to two decimal places; it is only the interpretation of them as evidence that is wrong.

The value $T_d = 1.73$ leads to the point estimate $\hat{\delta}_d = 0.454$ of $\delta_d = \mu_d/\sigma_d$. Further, using Equation (3.4), we obtain a 95 % confidence interval for δ_d of $[-0.06, 1.03]$.

4

Comparing treatment to control

4.1 Equal unknown precision

Data and model

- Given two independent sets of measurements: $x_1, \ldots, x_{n_1}$ on a control variable X, and $y_1, \ldots, y_{n_2}$ on a treatment variable Y, where the measurements on each variable have the same, but unknown, precision. The X-data are summarized in terms of the sample mean $\bar{x}$ and variance s_1^2, and similarly the pair $(\bar{y}, s_2^2)$ for the Y-data.

- The x_i's are regarded as independent observations which form a sample from a normal distribution with mean μ_1 and standard deviation σ; similarly for the y_j's, but the mean μ_2 could differ from μ_1 while the standard deviation is the same unknown σ.

- The *effect* is defined by $\theta = \mu_2 - \mu_1$ and the *standardized effect* by $d_{\text{Cohen}} = \theta/\sigma$. This standardized effect is often called Cohen's-d in the psychological literature (Cohen 1988) and *the effect size* in Hedges and Olkin (1985).

Questions

- What is the evidence for a treatment effect in a known direction? For example, what is the evidence against the null hypothesis $\mu_1 = \mu_2$ and for the alternative

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

$\mu_2 > \mu_1$? Or, in other words, is the treatment variable Y larger than the control variable X in the sense that $\mu_2 > \mu_1$?

- By symmetry, if alternatives $\mu_2 < \mu_1$ are of interest, the change of direction is reflected in the sign of the evidence T. Thus we only comment on the direction $\mu_2 > \mu_1$ and interpret positive evidence as support for $\mu_2 > \mu_1$.

Test statistic and distribution

- Now $\sigma^2\{1/n_1 + 1/n_2\}$ is the variance of $\bar{y} - \bar{x}$, the best estimator of $\mu_2 - \mu_1$. For $N = n_1 + n_2$ and $q = n_2/N$ it can be rewritten $\sigma^2/\{N(q(1-q)\}$. This variance, and hence the standard error of $\bar{y} - \bar{x}$, is clearly minimized for $q = 0.5$; that is, $n_1 = n_2$.

- A more appropriate standardized effect for differing sample sizes is $\delta{=}\sqrt{q(1{-}q)}\ d_{\text{Cohen}}$ of which an estimator is $\hat{\delta} = \sqrt{q(1-q)}\ (\bar{y} - \bar{x})/s_p$, where $s_p^2 = \{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\}/(n_1 + n_2 - 2)$ is the pooled estimate of the variance σ^2.

- The test statistic $t_{\text{pool}} = \sqrt{N}\,\hat{\delta}$. Large values of t_{pool} favor the alternative $\mu_2 > \mu_1$ over the null $\mu_2 = \mu_1$. The statistic t_{pool} has the noncentral $t_\nu(\lambda)$ distribution with $\nu = N - 2$ df , and noncentrality parameter $\lambda = \sqrt{N}\,\delta$ (see Johnson *et al.* (1995), p. 509).

Conversion to evidence, interpretation and confidence intervals

- Because the test statistic has the same noncentral t-distribution as in the one-sample problem of the last chapter, the transformation to evidence $T = T(t_{\text{pool}})$ is exactly the same here as it was there. Namely, $T = \sqrt{N}\,\mathcal{K}(\hat{\delta})$, where $\mathcal{K}$ is given by Equation (3.1) as $\mathcal{K}(\delta) = \sqrt{2}\ \sinh^{-1}(\delta/\sqrt{2}\,) = \sqrt{2}\ \ln(\delta/\sqrt{2} + \sqrt{1 + \delta^2/2}\,)$.

- Evidence is now centered on $\tau = \sqrt{N}\,\mathcal{K}(\sqrt{q(1-q)}\ d_{\text{Cohen}})$, where $\mathcal{K}$ is given by Equation (3.1). Clearly balanced sampling $q = 0.5$ is preferred, because it maximizes the expected evidence for fixed N.

- A 95 % confidence interval for d_{Cohen} is, for $c = z_{0.975} = 1.96$,

$$\left[\frac{1}{\sqrt{q(1-q)}}\,\mathcal{K}^{-1}\left(\frac{T-c}{\sqrt{N}}\right),\ \frac{1}{\sqrt{q(1-q)}}\,\mathcal{K}^{-1}\left(\frac{T+c}{\sqrt{N-1}}\right)\right], \tag{4.1}$$

where $\mathcal{K}^{-1}(y) = \sqrt{2}\ \sinh(y/\sqrt{2}\,) = \{\mathrm{e}^{y/\sqrt{2}} - \mathrm{e}^{-y/\sqrt{2}}\}/\sqrt{2}$ is the inverse function to $y = \mathcal{K}(\delta)$.

Choosing the sample size

- For expected evidence τ_1 when in fact $d_{\text{Cohen}} = d_1$ it suffices to take sample size
$$N_1 = \{\tau_1/\mathcal{K}(\sqrt{q(1-q)}\; d_1)\}^2. \tag{4.2}$$

4.2 Differing unknown precision

Data and model

- Given two independent sets of measurements: $x_1, \ldots, x_{n_1}$ on a control variable X, and $y_1, \ldots, y_{n_2}$ on a treatment variable Y, where the measurements on each variable have different unknown precision. The X-data are summarized in terms of the sample mean $\bar{x}$ and variance s_1^2, and similarly the pair $(\bar{y}, s_2^2)$ for the Y-data.
- The x_i's are regarded as independent observations which form a sample from a normal distribution with mean μ_1 and standard deviation σ_1; similarly y_i's are regarded as independent observations which form a sample from a normal distribution with mean μ_2 and standard deviation σ_2.
- The *effect* is defined by $\theta = \mu_2 - \mu_1$ and the *standardized effect* by $\delta = \theta/\sigma$, where σ is a scale parameter arising as follows. Let $N = n_1 + n_2$ and $\hat{\theta} = \bar{y} - \bar{x}$. Then define
$$\sigma^2 = N\,\mathrm{Var}[\hat{\theta}] = N\left\{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right\}. \tag{4.3}$$
Note that the standard error of $\hat{\theta}$ is $\mathrm{SE}[\hat{\theta}] = \sigma/\sqrt{N}$. For further discussion of this definition, see Section 21.2.

Questions

- Is the treatment variable Y larger/smaller than the control variable X in the sense that $\mu_2 > \mu_1$ or $\mu_2 < \mu_1$, respectively?
- These questions can be rewritten in terms of $\delta > 0$ and $\delta < 0$, where δ is the standardized effect defined by $\delta = (\mu_2 - \mu_1)/\sigma$.
- As before, we only comment on the direction $\mu_2 > \mu_1$.

Test statistic and distribution

- For $A = \sigma_1^2/n_1$, $B = \sigma_2^2/n_2$ define the Welch df by
$$\nu = (A + B)^2/\{A^2/(n_1 - 1) + B^2/(n_2 - 1). \tag{4.4}$$
Further define estimates $\hat{\nu}$, $\hat{\sigma}$ of ν, σ by substituting sample variances s_1^2, s_2^2 for the respective population variances σ_1^2, σ_2^2 in Equations (4.3), (4.4).

- Then the Welch statistic defined by $t_{\text{Welch}} = \sqrt{N}(\bar{Y}_{n_2} - \bar{X}_{n_1})/\hat{\sigma}$ has, under the null $\delta = 0$, an approximate Student t-distribution with $\hat{\nu}$ degrees of freedom.

Transformation to evidence

- The transformation of $t_{\text{Welch}} = \sqrt{N}(\bar{Y}_{n_2} - \bar{X}_{n_1})/\hat{\sigma}$ to the evidence scale $T = T(t_{\text{Welch}})$ is realized by Equation (21.6); it is similar to the *vst* of the Student t-distribution.

- For variance ratio $\varrho = \sigma_2^2/\sigma_1^2$ satisfying $0.5 \le \varrho \le 2$ and reasonably balanced sampling $0.5 \le n_2/n_1 \le 2$ the variance of T is stabilized near 1 and nominal 95 % confidence intervals for δ derived from T are reliable for all $|\delta| \le 1$, provided $N = n_1 + n_2 \ge 10$. These results improve with increasing sample sizes n_1, n_2; for further details see Section 21.4. In most applications $|\delta| \le 1$.

Interpretation

- The expected evidence $\tau = \sqrt{N}\,\mathcal{K}_\xi(\delta)$ is defined for each ξ, δ by $\mathcal{K}_\xi(\delta) = \sqrt{2/\xi}\ \sinh^{-1}(\delta\sqrt{\xi/2})$; where $\sinh^{-1}(x) = \ln(x + \sqrt{1+x^2})$. The parameter ξ is defined in Equation (21.7) and $\xi \approx N/\nu$; that is, ξ is roughly equal to the ratio of the total sample size N to Welch's df ν. Note that the expected evidence decreases in magnitude with increasing ξ, so it is desirable that ξ be near 1.

- The constant $\xi \ge 1$ and it can be shown that $\xi = 1$ when the sample sizes m, n are proportional to the standard deviations σ_1, σ_2, so if there is some knowledge of the ratio ϱ, the total sample size can be allocated accordingly. For example if $N = 30$ and one knows *a priori* that $\varrho \approx 4$ or $\sigma_2/\sigma_1 \approx 2$, then it is best to take $n_2/n_1 \approx 2$, that is, $n_1 = 10$ and $n_2 = 20$. Of course usually ϱ is unknown, and then balanced sampling $n_1 = n_2$ is recommended, for then $1 \le \xi \le 2$.

Choosing the sample size

- For balanced sampling $N = 2n_1$, the minimum value of $|\tau|$, as ξ varies, occurs for $\xi = 2$, and then $\tau = \sqrt{N}\,\mathcal{K}_2(\delta) = \sqrt{N}\ \sinh^{-1}(\delta)$. Therefore the minimum sample size N_1 required to guarantee expected evidence $\tau_1 = \sqrt{N}\ \sinh^{-1}(\delta)$ when $\delta = \delta_1$ is

$$N_1 = \{\tau_1/\sinh^{-1}(\delta_1)\}^2. \tag{4.5}$$

  For example, to guarantee 'moderate' expected evidence $\tau_1 = 3.3$ for $\delta > 0$ when in fact $\delta = \delta_1 = 0.5$ one needs $N_1 = 47$, or equal sample sizes of 24 each. For only 'weak' expected evidence of 1.645 under the same conditions one needs equal sample sizes of 6 each.

- For unbalanced sampling, use Equation (21.9).

Confidence intervals

- An approximate $100(1-\alpha)\,\%$ confidence interval for θ is given by the Welch t-interval
$$\left[\hat{\theta} - t_{\hat{\nu},1-\alpha/2}\,\frac{\hat{\sigma}}{\sqrt{N}}\,,\ \hat{\theta} + t_{\hat{\nu},1-\alpha/2}\,\frac{\hat{\sigma}}{\sqrt{N}}\right]. \tag{4.6}$$

- For $c = z_{0.975}$ an approximate 95 % confidence interval for δ based on the one-sided evidence $T = T(t_{\text{Welch}})$ is given by
$$\left[\sqrt{\frac{2}{\hat{\xi}}}\,\sinh\left\{\sqrt{\frac{\hat{\xi}}{2N}}\,(T-c)\right\}\,,\ \sqrt{\frac{2}{\hat{\xi}}}\,\sinh\left\{\sqrt{\frac{\hat{\xi}}{2N}}\,(T+c)\right\}\right], \tag{4.7}$$
where ξ is estimated by $\hat{\xi} = N/\hat{\nu}$.

4.3 Examples

In the first example the assumption of equal unknown precision appears reasonable, so the methods of Section 4.1 are employed, while in the second example unequal precision is apparent and so the methods of Section 4.2 are illustrated.

4.3.1 Drop in systolic blood pressure

Summary statistics from seven studies in the review by Mulrow *et al.* (2004) are shown in Table 4.1. In each study the sample mean $\bar{y}_k$ gives the average drop in systolic blood pressure for a group of patients following a weight reducing diet, and $\bar{x}_k$ is the average drop for a control group. For every one of the studies $s_{1k} \approx s_{2k}$, so the pooled estimate $s_{\text{pool},\,k}$ of a common unknown standard deviation σ_k is computed. The two-sample pooled t-statistic with $\nu_k = n_{1k} + n_{2k} - 2$ degrees of freedom and

Table 4.1 Seven studies comparing drop in systolic blood pressure for treated patients undergoing a weight-loss regime (summarized by n_2, $\bar{y}$, s_2) with control patients not undergoing a weight-loss regime (summarized by n_1, $\bar{x}$, s_1). The estimated effect $\hat{\theta}_k$, pooled sample standard deviation $s_{\text{pool},\,k}$, two-sample t-statistic $t_{\text{pool},\,k}$ and evidence for a positive effect T_k for each k are also tabled.

k	n_{1k}	$\bar{x}_k$	s_{1k}	n_{2k}	$\bar{y}_k$	s_{2k}	N_k	$\hat{\theta}_k$	$s_{\text{pool},\,k}$	$t_{\text{pool},\,k}$	T_k
1	24	0.2	13.8	27	−4.8	13.8	51	−5.0	13.80	−1.29	−1.24
2	18	7.4	8.1	20	13.3	8.1	38	5.9	8.10	2.24	2.11
3	64	4.0	15.7	66	11.0	17.1	130	7.0	16.43	2.43	2.39
4	9	−3.0	13.5	10	4.0	15.3	19	7.0	14.48	1.05	0.94
5	25	15.0	16.5	24	8.0	20.4	49	−7.0	18.51	−1.32	−1.27
6	5	2.5	5.1	5	9.8	7.1	10	7.3	6.18	1.87	1.42
7	14	9.9	6.4	19	12.5	6.3	33	2.6	6.34	1.16	1.09

Table 4.2 Statistical summaries of eight studies from Mumford *et al.* (1984) are listed in columns 1–7. The results compare length of stay in hospital for patients receiving psychotherapy summarized by $(n_2, \bar{y}, s_2)$ and control groups with length of stay data summarized by $(n_1, \bar{x}, s_1)$. For each study k are also given the Welch degrees of freedom $\hat{\nu}_k$, the estimated effect $\hat{\theta}_k$, the estimated scale parameter $\hat{\sigma}_k$ and the 95 % Welch confidence interval (4.6) for the unknown θ_k.

k	n_{1k}	$\bar{x}_k$	s_{1k}	n_{2k}	$\bar{y}_k$	s_{2k}	$\hat{\nu}_k$	$\hat{\theta}_k$	$\hat{\sigma}_k$	$[L_k,\ U_k]$
1	13	6.50	3.80	13	5.00	4.70	22.99	−1.50	8.55	[−4.98, +1.98]
2	50	6.10	2.30	30	4.90	1.71	74.31	−1.20	4.03	[−2.10, −0.30]
3	35	24.90	10.65	35	22.50	3.44	41.02	−2.40	15.83	[−6.22, +1.42]
4	20	12.30	1.66	20	12.50	1.47	37.45	0.20	3.14	[−0.80, +1.20]
5	10	3.19	0.79	10	3.37	0.92	17.60	0.18	1.72	[−0.63, +0.99]
6	14	5.50	0.90	13	4.90	1.10	23.26	−0.60	2.02	[−1.40, +0.20]
7	9	12.78	2.05	9	10.56	1.13	12.45	−2.22	3.31	[−3.92, −0.52]
8	8	7.38	1.41	8	6.50	0.76	10.75	−0.88	2.26	[−2.14, +0.38]

the evidence for a positive effect T_k are also shown in Table 4.1. It is apparent that only studies 2 and 3 would reject the null hypothesis of no effect at level 0.05, and the evidence for a positive effect, shown in the last column, is only weak for these two studies. A 95 % confidence interval for Cohen's-d in the second study, based on Equation (4.1), is [0.05, 1.39]. Later in Chapter 25 we will demonstrate how one can combine the evidence in the seven studies.

4.3.2 Effect of psychotherapy on hospital length of stay

Mumford *et al.* (1984) compare the effectiveness of treatment 'psychotherapy' with control 'no therapy' for reducing length of stay in hospital in days for eight different studies. The data are shown in Table 4.2. The sample variances indicate that heteroscedasticity is present within most studies, so the Welch type t-statistic is appropriate.

In the first study the estimated effect $\theta_1 = \bar{x}_1 - \bar{y}_1 = -1.5$ days, the standard error of this estimate is $\hat{\sigma}_1/\sqrt{N_1} = 8.548/\sqrt{26} = 1.676$. The 95 % Welch confidence interval for θ_1 is $[-4.98, +1.98]$, which contains 0, so the hypothesis $\theta_1 = 0$ would not be rejected at level 0.05. By the same argument, it is clear that only studies 2 and 5 would reject at level 0.05 the null hypothesis of psychotherapy having no effect. These results can be obtained on most statistical packages. But if one wants to compare the studies, it is better to look at standardized effects, rather than raw effects, because then the studies are all compared on the basis of a scale-free measurement.

In Table 4.3 are shown the results for the standardized effect analysis. The estimated standardized effect in the first study is $\hat{\delta}_1 = \hat{\theta}_1/\hat{\sigma}_1 = -0.175$ and a 95 % confidence interval for δ_1 is $[-0.570, +0.218]$. Because this interval does contain 0 we could not reject $\delta_1 = 0$ at level 0.05, confirming the small magnitude of the Welch statistic $t = -0.895$. Only studies 2 and 7 provide level 0.05 significance that psychotherapy reduces length of stay. This was already found in Table 4.2. However, now we can actually see how much evidence there is for a positive effect in each study.

Table 4.3 For each study k in the Mumford *et al.* (1984) review are listed $\hat{\xi}_k \approx N_k/\hat{\nu}_k$, the Welch statistic $t_{\text{Welch},\,k}$ and the evidence T_k for a positive standardized effect δ_k which lies in it. In the last column are the 95 % confidence intervals based on (4.7) for the unknown values of δ_k.

k	$\hat{\xi}_k$	$\hat{\delta}_k$	$t_{\text{Welch},k}$	T_k	$[L_k, U_k]$
1	1.24	−0.175	−0.895	−0.86	[−0.57, +0.22]
2	1.11	−0.298	−2.662	−2.61	[−0.52, −0.07]
3	1.79	−0.152	−1.269	−1.24	[−0.39, +0.09]
4	1.13	0.064	0.403	0.39	[−0.25, +0.38]
5	1.28	0.105	0.469	0.44	[−0.34, +0.55]
6	1.27	−0.297	−1.544	−1.47	[−0.69, +0.09]
7	1.72	−0.671	−2.845	−2.53	[−1.24, −0.14]
8	1.83	−0.388	−1.554	−1.41	[−0.94, +0.14]

Note that the alternative $\delta < 0$ is of interest here. The positive evidence for negative effect is equivalent to negative evidence for positive effect found here. Six of the eight studies yield standardized effects which suggest by direction that psychotherapy does make a difference and these effects are all much greater in magnitude than the other two. Thus it is plausible that there is at least weak evidence for an overall negative effect in the eight studies, i.e. the overall reduction in length of stay.

5

Comparing *K* treatments

In this chapter we consider the simplest case of treatment comparisons. Based on K samples taken under K different conditions we want to know whether the conditions lead to notable changes in the sample means.

5.1 Methodology

Data and model

- We are given K sequences of measurements of some outcome variable Y: $y_{11}, \ldots, y_{1n_1}$ through $y_{K1}, \ldots, y_{Kn_K}$. The measurements are taken under varying conditions, either by applying different treatments or by modifying in some other way the circumstances of the measurements. Following tradition, we call these circumstances ‘treatments’.
- The measurements are modelled as samples from K normal populations with means μ_k for the kth sample and with equal and unknown variance σ^2.

Questions

- What is the evidence for differential effects of the K conditions? The various means μ_k are called treatment effects and the null situation occurs when they are all equal; that is, $H_0 : \mu_1 = \cdots = \mu_K$. The alternative we are interested in simply states that H_0 is untrue.
- Because our alternative does not describe a precise deviation from the null situation, no direction or sidedness is involved.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte
© 2008 John Wiley & Sons, Ltd

Test statistic and distribution

- The treatment effects can be estimated by the sample means $\hat{\mu}_k = (y_{k1} + \cdots + y_{kn_k})/n_k = \bar{y}_k$. The total sample size is $N = n_1 + \cdots + n_K$ and the mean of all measurements is $\bar{y} = (n_1\bar{y}_1 + \cdots + n_K\bar{y}_K)/N$. A large variance of the sample means $[n_1\,(\bar{y}_1 - \bar{y})^2 + \cdots + n_K\,(\bar{y}_K - \bar{y})^2]/(K-1)$ supports the alternative of unequal treatment effects, but this statistic is difficult to calibrate because its size also depends on the size of the variance σ^2.

- The sample variances for the kth sample is $s_k^2 = [(y_{k1} - \bar{y}_k)^2 + \cdots + (y_{kn_k} - \bar{y}_k)^2]/(n_k - 1)$. These can be pooled to create a stronger estimate $[(n_1 - 1)s_1^2 + \cdots + (n_K - 1)s_K^2]/(N - K)$. The rescaled test statistic is

$$S = \frac{[n_1\,(\bar{y}_1 - \bar{y})^2 + \cdots + n_K\,(\bar{y}_K - \bar{y})^2]/(K-1)}{[(n_1 - 1)s_1^2 + \cdots + (n_K - 1)s_K^2]/(N-K)}.$$

- Under the null hypothesis the test statistic S has an F-distribution $F_{K-1,N-K}$ with $K-1$ and $N-K$ degrees of freedom and the usual procedure consists in deriving a p-value from a tabulated F-distribution.

- Under alternatives S has a noncentral F-distribution $ncF_{K-1,N-K}(\lambda)$ with noncentrality parameter $\lambda = [n_1(\mu_1 - \mu)^2 + \cdots + n_K(\mu_K - \mu)^2]/\sigma^2$.

Transformation to evidence

- To convert the test value into evidence, we make use of the inverse of the hyperbolic cosine function

$$\cosh^{-1}(y) = \ln(y + \sqrt{y^2 - 1}).$$

Furthermore, let $m = F_{K-1,N-K}^{-1}(0.5)$ denote the median (50 % quantile) of the F-distribution with $K-1$ and $N-K$ degrees of freedom. For the computation of the evidence we make a distinction between large and small values of the test statistic. For values of S exceeding the median, the evidence is

$$T = \sqrt{\frac{N-K}{2}}\left(\cosh^{-1}\left(\frac{(K-1)\,S + N - K}{\sqrt{(N-K)\,((K-1)m + N - K)}}\right) - \cosh^{-1}\left(\sqrt{\frac{(K-1)m + N - K}{N-K}}\right)\right).$$

For values of S below the median m, essentially the same formula can be used. First we compute the flipped value of the test statistic

$$S^* = F_{K-1,N-K}^{-1}\,(1 - F_{K-1,N-K}(S)).$$

Then we compute the evidence with changed sign

$$T = -\sqrt{\frac{N-K}{2}}\left(\cosh^{-1}\left(\frac{(K-1)\,S^{*}+N-K}{\sqrt{(N-K)\,((K-1)m+N-K)}}\right)\right.$$
$$\left. - \cosh^{-1}\left(\sqrt{\frac{(K-1)m+N-K}{N-K}}\right)\right).$$

For additional explanations and derivations, the reader should consult Chapter 23.

Interpretation

- The crucial quantity for the test statistic S is the noncentrality parameter λ. The bigger this value, the further we are from the null hypothesis. A standardized effect for the $F_{K-1,N-K}(\lambda)$-distribution is often defined as λ or λ/N.
- The evidence T is calibrated and can be interpreted on the scale of a normal distribution with variance 1. For the evidence statistic we have $T \sim N(\sqrt{N}\mathcal{K}, 1)$, where $\mathcal{K} = \mathcal{K}(\lambda)$ depends on N, K and λ.
- The value of $\mathcal{K}$ is the transformed effect computed as follows:

  $$\mathcal{K}(\lambda) = \sqrt{\frac{N-K}{2N}}\left(\cosh^{-1}\left(\frac{(K-1)m+\lambda+N-K}{\sqrt{(N-K)\,((K-1)m+N-K)}}\right)\right.$$
  $$\left. - \cosh^{-1}\left(\sqrt{\frac{(K-1)m+N-K}{N-K}}\right)\right).$$

  If $\mathcal{K}$ exceeds zero and as N increases, the evidence in favor of the alternatives will increase.
- As in all the other tests discussed in this book, the key inferential function translates the apparent effect λ into a statistically meaningful transformed effect. The transformed effect is estimated by $\hat{\kappa} = T/\sqrt{N}$.

Choosing the sample size

- For a known or assumed value of the noncentrality parameter λ one can choose the sample size N necessary to reach any desired expected evidence τ. The equation to solve is $\tau = \sqrt{N}\mathcal{K}(\lambda)$, which is not as simple as in some of the other cases, since the Key Inferential Function itself depends on N and K. We recommend to solve the equation by trial-and-error.
- Instead of fixing the expected amount of evidence τ, we may be interested in designing the study to have a certain level (probability of false rejection) α and power (probability of true rejection) $1-\beta$. In this case we need to solve $\sqrt{N}\mathcal{K}(\lambda) = z_{1-\alpha} + z_{1-\beta}$, where $z_p = \Phi^{-1}(p)$ is the p quantile of the standard normal distribution.

Confidence intervals

- Let $c = z_{0.975} = 1.96$. The interval $T \pm 1.96$ is a confidence interval for $\sqrt{N}\,\mathcal{K}(\lambda)$.
- A nominal 95 % confidence interval for λ is given by

$$[L, U] = [\mathcal{K}^{-1}(T - c), \mathcal{K}^{-1}(T + c)],$$

where

$$\mathcal{K}^{-1}(y) = \cosh\left(y\sqrt{2/(N-K)} + \cosh^{-1}\left(\sqrt{\frac{Km + (N-K)}{(N-K)}}\right)\right) \\ \times\sqrt{(N-K)\,(Km + (N-K))} - (Km + (N-K)).$$

is the inverse function to $y = \mathcal{K}(\lambda)$.

5.2 Examples

Comparing a treatment to a control is a common practice in many applications. This comparison can be based on two series of measurements, one of which under the condition of treatment and the other under the control condition. Alternatively, one may form a sample of matched pairs and apply the treatment and the control to one member of each of the pairs. The generalization to K treatments with $K > 2$ is an equally useful method. Such experiments may again be performed in the form of independent series of measurements or in the form of K blocks of size n with random allocation of the treatments to the units of each block (randomized block design).

The test we discuss here does not involve pairwise comparisions or other methods to determine the precise differences between the treatment effects. We are only concerned with the evidence in favor of the alternative that 'some' difference between the treatment effects exists.

5.2.1 Characteristics of antibiotics

Ziv and Sulman (1972; cited in Larsen and Marx (1986), p. 504) gave measurements of Y = binding percentage characteristics of five antibiotics. Table 5.1 contains the data and Figure 5.1 shows a plot of the five samples.

The F-test statistic is $S = (1480.8/(K-1))/(135.8/(N-K)) = 40.9$, which has to be compared to an F-distribution $F_{4,15}$ with a 50 % quantile of 0.88 and a 95 % quantile of 3.06. The evidence statistic is $T = 7.1$. Both computations lead to identical conclusions. The evidence is seven standard deviations from zero, which means that the evidence against the null hypothesis is very strong. The p-value is 6×10^{-8} which would also be interpreted as a very strong indication in favor of the alternative.

Table 5.1 Binding percentages of five antibiotics. This is an example with $K = 5$ different conditions. The sample size $n = 4$ is the same in each case, which leads to a total size of $N = Kn = 20$.

Penicillin G	Tetracycline	Streptomycin	Erythromycin	Chloramphenicol
29.6	27.3	5.8	21.6	29.2
24.3	32.6	6.2	17.4	32.8
28.5	30.8	11.0	18.3	25.0
32.0	34.8	8.3	19.0	24.2
Averages				
28.6	31.4	7.8	19.1	27.8
Sample variances				
10.4	10.1	5.7	3.3	15.9

5.2.2 Red cell folate levels

Amess *et al.* (1978; discussed as Example 9.8.2 in Altman (1991)) contains measurements of the outcome variable $Y =$ red cell folate levels in patients treated with nitrous oxide, N_2O (often called laughing gas). This is usually administered in a 50/50 mixture with oxygen as a simple anesthetic agent. Among the toxicological side effects is the inhibition of an enzyme, leading to impairment of folate metabolism. In this example,

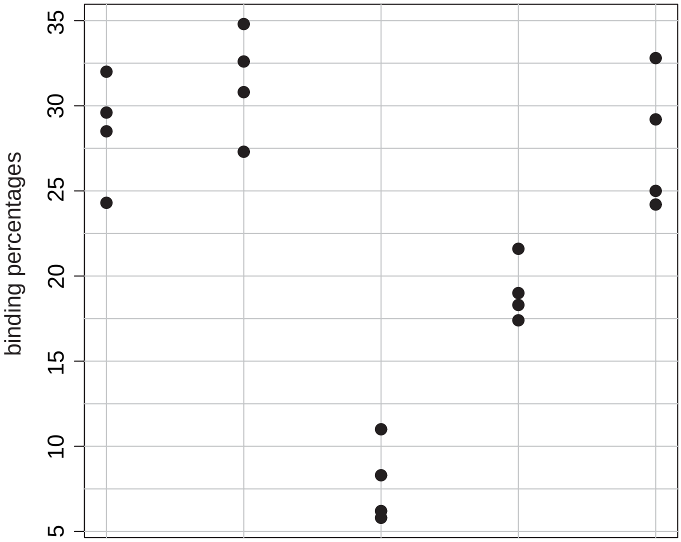


Figure 5.1 The five samples are shown alongside each other on a common vertical axis. The order of the samples is the same as in Table 5.1.

Table 5.2 Red cell folate levels (μg/l) in three groups of patients given different concentrations of nitrous oxide-enriched ventilations. This is an example with $K = 3$ different conditions. The sample sizes are $n_1 = 8$, $n_2 = 9$ and $n_3 = 5$. The total sample size is $N = 22$.

Long	Short	Oxygen only
243	206	241
251	210	258
275	226	270
291	249	293
347	255	328
354	273	
380	285	
392	295	
	309	
Averages		
316.6	256.4	278.0
Sample variances		
3447.7	1378.0	1139.5

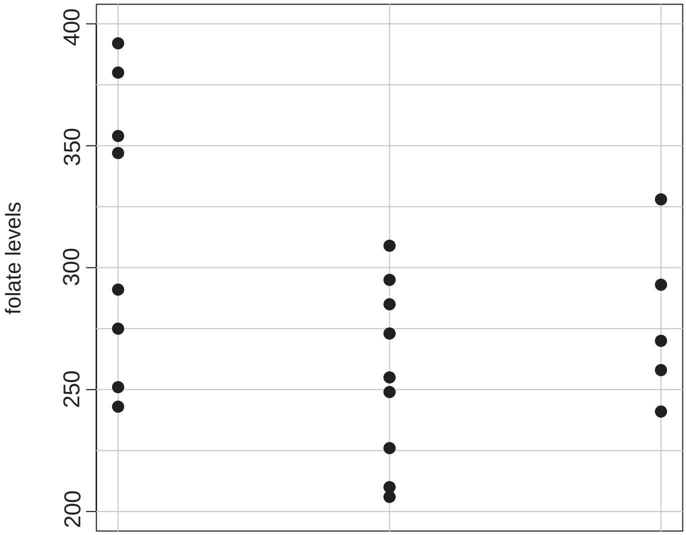


Figure 5.2 The three samples are shown alongside each other on a common vertical axis. The order of the samples is the same as in Table 5.2.

three groups of patients undergoing cardiac bypass surgery are considered. The treatments given to these patients were distinguished according to the mixture they breathed and the duration of the ventilation regime. Table 5.2 contains the data and Figure 5.2 shows a plot of the three samples.

The F-test statistic is $S = (15515.8/(K-1))/(39716.1/(N-K)) = 3.71$, which has to be compared to an F-distribution $F_{2,19}$ with a 50 % quantile of 0.72 and a 95 % quantile of 3.52. The evidence statistic is $T = 1.64$, which speaks weakly in favor of the alternative of some difference between the treatments. The p-value is equal to 0.044, close to the traditional 5 %.

The expected unit evidence is $\hat{\mathcal{K}} = 1.64/\sqrt{N} = 0.35$ with an associated confidence interval $[0.35 \pm 0.21]$. By how much would we have to increase the study size N in order to reach moderate evidence of 3.3? Using $\hat{\mathcal{K}} = 0.35$, we must solve $\sqrt{N_{\text{study}}} = 3.3/0.35 = 9.4$, which gives $N_{\text{study}} \approx 88$. We would thus have to quadruple the size of the study in order to reach moderate evidence. Of course, this prediction is subject to considerable uncertainty. Had we used the lower confidence point for $\hat{\mathcal{K}}$, we would have obtained $N_{\text{study}} \approx 555$.

6

Evaluating risks

In prospective studies the risk or incidence of contracting a disease is often represented by a probability p, the probability that someone drawn from a cohort contracts a disease during a certain period of time. Or p could represent the prevalence of a disease within a certain population at the present time. In either case a random sample of individuals is examined at a fixed time and the number within the sample with the disease is noted. The question is then how to use this information to estimate p, or to test hypotheses regarding p.

6.1 Methodology

Data and model

- The data are a set of n dichotomous observations; that is, each taking on one of two possibilities, say D for diseased, $\bar{D}$ for not diseased, or labeled numerically by 1 and 0.
- The binomial model assumes that there are n independent, identically distributed variables, say $I_1, \ldots, I_n$, with $P(I_j = 1) = p$, $P(I_j = 0) = 1 - p$, where $0 < p < 1$ is unknown.

Questions

- What is the evidence that the risk p exceeds a certain fixed level p_0?
- What is a confidence interval for the risk p?

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

Test statistic and distribution

- The test statistic is the number of 1's amongst the n outcomes, $S = \sum_{j=1}^{n} I_j$; it has the binomial distribution with parameters n, p.

Transformation to evidence and distributional properties

- Let $\tilde{p} = (S + 3/8)/(n + 3/4)$. Then the evidence for the alternative $p > p_0$ to the null $p = p_0$ is given by the classic transformation
$$T = 2\sqrt{n}\,\left\{\arcsin(\sqrt{\tilde{p}}\,) - \arcsin(\sqrt{p_0}\,)\right\}.$$
- This T is approximately normal for $np(1 - p) \geq 5$.
- The expected evidence $E[T] \doteq \sqrt{n}\,\mathcal{K}(p)$, where the Key Inferential Function is defined by
$$\mathcal{K}(p) = 2\left\{\arcsin(\sqrt{p}\,) - \arcsin(\sqrt{p_0}\,)\right\}.$$
- The evidence T has standard deviation lying between 0.95 and 1.0 for $0.2 < p < 0.8$ for sample size $n = 9$, and this range expands to $0.07 < p < 0.93$ for $n = 30$. For any n, as p approaches 0 or 1, the standard deviation of T approaches 0; but this does not mean that T is not a good estimator of its expected value. For more information, see Figure 18.1 and accompanying text.

Interpretation

- Positive values of T are evidence for the alternative $p > p_0$, while the magnitude $|T|$ of a negative value of T is positive evidence for the alternative $p < p_0$. Evidence $T^{\pm}$ for the two-sided alternative $p \neq p_0$ can be obtained from $|T|$ via the transformation (2.3).

Choosing the sample size

- For testing $p = p_0$ against $p > p_0$ one may choose n_1 so that the expected evidence for a fixed p_1 of interest is at least τ_1. This requires n_1 to satisfy $\tau_1 \leq \sqrt{n_1}\,\mathcal{K}(p_1)$, or $n_1 \geq \{\tau_1/\mathcal{K}(p_1)\}^2$.
- For example, if the null is $p = 0.5$ to achieve 'strong' expected evidence $\tau_1 = 5$ against $p = 0.9$ one requires $n_1 \approx 29$. Some other values are also shown in Table 6.1.

Confidence intervals

- Letting $\tilde{p} = (S + 3/8)/(n + 3/4)$ and $T = 2\sqrt{n}\,\arcsin\left(\sqrt{\tilde{p}}\,\right)$, a 95 % confidence interval for p is given by
$$\left[\,\left\{\sin\left(\frac{T - z_{0.975}}{2\sqrt{n}}\right)\right\}^2,\ \left\{\sin\left(\frac{T + z_{0.975}}{2\sqrt{n}}\right)\right\}^2\,\right].$$
It is understood that if the sine values are less than 0 or greater than 1, they are replaced, respectively, by 0 and 1, before squaring.

Table 6.1 Approximate sample sizes required to achieve weak, moderate or strong expected evidence for alternatives p_1 to the null $p_0 = 0.5$.

p_1	$\arcsin(\sqrt{p_1})$	$\tau_1 = 1.645$	$\tau_1 = 3.3$	$\tau_1 = 5.0$
0.5	0.78540	—	—	—
0.6	0.88608	67	267	617
0.7	0.99116	16	64	148
0.8	1.10715	7	26	60
0.9	1.24905	4	13	29

- These intervals are much more accurate than traditional large sample intervals of the form $\hat{p} \pm z_{0.975}\sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p} = X/n$ (see Section 18.2).
- When p is near 0, confidence intervals for p are often derived after a log-transformation of $\hat{p} = S/n$. Such intervals are comparable in performance to those based on the formula displayed above (see Section 18.4). A rule of thumb suggested based on simulations reported in Section 18.4 is that the when conditions $np(1-p) \geq 5$ and $n \geq 25$ are satisfied, then the arcsine intervals will have empirical coverage between 93 and 97 %; and for $np(1-p) \geq 11$ and $n \geq 100$, the coverages will lie between 94 and 96 %.

6.2 Examples

These methods have already been illustrated for the case of $p_0 = 0.5$ in matched pair experiments in Section 1.2.

6.2.1 Ultrasound and left-handedness

A study by Salvesen *et al.* (1993) found a slight positive association between *in utero* routine ultrasonography and subsequent left-handedness of 8- and 9-year-old children. Similar reports for only boys in a different study were reported by Kieler *et al.* (1998). If the proportion of left-handers in the general population is $p_0 = 0.1$, how large a sample is required to obtain strong evidence that *in utero* routine ultrasonography leads to a proportion p of left-handers which exceeds the general population proportion 0.1 by 10 %? That is, what is the minimum sample size required to obtain expected evidence 5 for an alternative $p = p_1 = 0.11$?

We require $n_1 \geq \{\tau_1/\mathcal{K}(p_1)\}^2 = (5/0.03263)^2 = 23\,481.3$, or 23 482. For only moderate evidence 3.3 of a 10 % increase, one needs a minimum sample size of $n_1 = 10\,229$.

6.2.2 Treatment of recurrent urinary tract infections

If untreated, recurrent urinary tract infections continue in 65 % of observed patients (see Section 19.5). Let p represent the risk of continued infection following treatment by antibiotics. In study 2 of Table 19.1 eight of 21 patients treated by antibiotics had further infections during the study period. How much evidence is there for the alternative $p < 0.65$ to the null $p = 0.65$ based on these data?

In the notation of this chapter $n = 21$ and $S = 8$. An estimate of p is $\tilde{p} = (S + 3/8)/(n + 3/4) = 0.3895$. Hence the evidence for the alternative $p < 0.65$ is $T = 2\sqrt{n}\,\{\arcsin(\sqrt{p_0}\,) - \arcsin(\sqrt{\tilde{p}}\,)\} = 2\sqrt{21}\,\{\arcsin(\sqrt{0.65}\,) - \arcsin(\sqrt{0.3895}\,)\} = 2.4$, which is between weak and moderate. An analysis based on comparing treatment patients to similar controls is given in Section 7.2.1.

7

Comparing risks

Unknown probabilities of a binary outcome (of survival, contracting a disease, say) for individuals in two groups, treatment and control, are often called risks. The ‘treatment’ could be exposure to a risk factor, drug intervention, surgery, etc. The risk p_1 to an individual in the control group is compared to the risk p_2 in a treated group after obtaining independent estimates of these parameters. Here we present new methods for inference on the risk difference $p_1 - p_2$, the relative risk p_1/p_2 and the odds ratio $p_1(1 - p_2)/\{p_2(1 - p_1)\}$. It turns out that the evidence for a higher risk in one of the groups is the same regardless of how this difference is measured: by the risk difference, or relative risk, or odds ratio. We start from estimating the risk difference Δ which has an advantage of linearity in probabilities. Standard methods for the relative risk and odds ratio can be found in Chapter 19.

7.1 Methodology

Data and model

- There are n_1 subjects from a population of control subjects having proportion p_1 at risk; and x_1 of the n_1 are found to have the binary outcome of interest, say disease. Similarly x_2 of the n_2 subjects from a population of treated subjects having proportion p_2 at risk have the outcome of interest.

- Given $X_1,\ X_2$ independent, with each X_i having the binomial distribution with parameters (n_i, p_i) for some $0 < p_i < 1$.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte
© 2008 John Wiley & Sons, Ltd

Questions

- What is the evidence for treated subjects having a lower risk than control subjects? That is, for $\Delta = p_1 - p_2 > 0$; or, equivalently, for the relative risk $\text{RR} = p_1/p_2 > 1$ or for the odds ratio $\text{OR} = p_1(1-p_2)/\{p_2(1-p_1)\} > 1$?
- What are confidence intervals for the risk difference, relative risk and the odds ratio?

Test statistic and distribution

- The test statistic and estimator of Δ is defined by $S = \hat{\Delta} = \tilde{p}_1 - \tilde{p}_2$, where $\tilde{p}_1 = (X_1 + 0.5)/(n_1 + 1)$, $\tilde{p}_2 = (X_2 + 0.5)/(n_2 + 1)$. Its distribution is complicated and the standardized version $(\hat{\Delta} - 0)/\text{SE}[\hat{\Delta}]$ converges to the standard normal distribution much slower than commonly believed.

Transformation to evidence

- Let $N = n_1 + n_2$ be the total sample size, $q = n_2/N$ the proportion of the total allotted to the second sample. Define $p = qp_1 + (1-q)p_2$ and $\zeta = p(1-p)/\{q(1-q)\}$. Substitute $\tilde{p}_1, \tilde{p}_2$ for p_1, p_2 in the formula for ζ to obtain an estimator $\tilde{\zeta}$ of it. Then the evidence for the alternative $\Delta > 0$ to the null $\Delta = 0$ is defined by
$$T = \sqrt{N}\ \arcsin\left(S/\sqrt{\tilde{\zeta}}\ \right). \tag{7.1}$$
- It is shown in Chapter 19 that T is approximately normal for a wide range of parameters. Further $\tau = \text{E}[T] \doteq \sqrt{N}\,\mathcal{K}(\rho)$, where $\rho = \Delta/\sqrt{\zeta}$, is the *correlation effect size* introduced in Section 1.3.1, and the Key Inferential Function is simply $\mathcal{K}(\rho) = \arcsin(\rho)$.

Interpretation

- The *standardized effect* for the difference Δ is $\delta = \Delta/\sqrt{N\,\text{Var}[\hat{\Delta}]} = \Delta/\sqrt{\zeta - \Delta^2}$. The Key can be reexpressed as a function of the standardized effect using the fact that $\rho = \delta/\sqrt{1+\delta^2}$.
- To uniformly maximize the magnitude of the expected evidence $|\text{E}[T]|$ by choice of sample size allocation q, it suffices to uniformly minimize $\zeta = \{p(1-p)\}/\{q(1-q)\}$ by choice of q. Now for any q, the numerator of ζ has maximum value 0.25, and thus the maximum value of ζ over the parameter space can be minimized by maximizing the denominator of ζ; that is, by choosing $q = 0.5$, or $n_1 = n_2$.

Choosing the sample size

- In order to attain expected evidence $\tau_1 \doteq \sqrt{N}\ \arcsin(\rho_1)$ for a correlation effect size ρ_1 one requires $N \geq \{\tau_1/\arcsin(\rho_1)\}^2$. In particular, to attain ‘moderate’ expected evidence of $3.3 = 2 \times 1.645$ for $\rho_1 = 0.5$, one needs a total sample size of $N \geq (6 \times 3.3/\pi)^2 = 39.7$, or $N = 40$. This calculation masks the fact that depending on ζ, that is, on the ratio of sample sizes and resulting value of p, very different raw effects Δ may result in the same value of ρ, and therefore the same evidence.

- To achieve an expected evidence of $\tau_1 \doteq \sqrt{N}\ \arcsin(\Delta_1/\sqrt{\zeta}\,)$ against an effect Δ_1, it suffices, for any fixed q, to take

$$N \geq \{\tau_1/\arcsin(\,2\sqrt{q(1-q)}\ \Delta_1)\,\}^2.$$

In particular, for $\tau_1 = 3.3$ and $\Delta_1 = 0.5$, it suffices to take equal sample sizes totalling $N = 40$. For unequal sample sizes, a larger total is required.
Often a baseline value for the control risk is known, and thus when the desired risk difference is specified, so is p, and the required sample size N is determined.

Confidence intervals

- A nominal 95 % confidence interval for ρ is given by

$$[L, U] = \left[\sin\left(\frac{T - z_{0.975}}{\sqrt{N}}\right),\ \sin\left(\frac{T + z_{0.975}}{\sqrt{N}}\right)\right].$$

- And this leads immediately to intervals having the same confidence for the standardized effect $\delta = \rho/\sqrt{1-\rho^2}$, namely

$$\left[\frac{L}{\sqrt{1-L^2}},\ \frac{U}{\sqrt{1-U^2}}\right].$$

- Further, the above 95 % confidence intervals $[L, U]$ for $\rho = \Delta/\sqrt{\zeta}$ can be multiplied by $\sqrt{\tilde{\zeta}}$, where $\tilde{\zeta} = \{\tilde{p}(1-\tilde{p})\}/\{q(1-q)\}$ is an estimate of ζ, to yield nominal 95 % confidence intervals $[L_\Delta, U_\Delta]$ for Δ. Despite the extra estimate involved, these intervals tend to have better coverage than the corresponding intervals for ρ.

- The above preservation of intervals under transformations tacitly assumed that the argument of the sine function in the definition of L, U lies within the interval $[-\pi/2,\ \pi/2]$ (wherein the sine function is strictly monotonic increasing). This requires $|T| < 3.0787\sqrt{N}$. Even for sample sizes as small as $n_1 = n_2 = 8$, this means $|T| < 12.3$. Evidence T with magnitude 5 is considered ‘strong’, so this restriction on T is likely to be met in applications.

- Simulation studies of the empirical coverage probabilities are available in Sections 19.2 and 19.3. The empirical coverage of intervals for ρ for balanced sampling with $n_1 = n_2 = 9$ and $p = 0.5$ ranges from 94.5 to 98 % for all ρ not too near 1. But when $p = 0.2$ much larger sample sizes are required to achieve the same coverage. The corresponding intervals for Δ have more accurate coverage.

Extensions to relative risk and odds ratio

- Using the identities $p_1 = p + (1 - q)\Delta$ and $p_2 = p - q\Delta$ one may rewrite the relative risk in terms of p and Δ and similarly for the odds ratio. From these expressions one can see that for fixed p, both the RR and OR are strictly increasing in Δ. Since the evidence $T = \sqrt{N}\ \arcsin\bigl(S\tilde{\zeta}^{-1/2}\bigr)$ has been derived in Chapter 19 by a conditional argument, given $\tilde{p} = p$, this evidence for $\Delta > 0$ can serve as evidence for RR > 0 or for OR > 0.

- In view of the above remarks, a confidence interval $[L, U]$ for Δ can be transformed into one for the relative risk by substituting the endpoints L, U for Δ in the expression $\text{RR} = (\tilde{p} + (1 - q)\Delta)/(\tilde{p} - q\Delta)$ to obtain the endpoints of an interval for the relative risk. Coverage of these confidence intervals, unlike the coverage for confidence intervals for Δ given above, has not yet been investigated by simulations. Until this has been done, we recommend instead the standard confidence intervals for RR or OR given in Section 19.4.

7.2 Examples

7.2.1 Treatment of recurrent urinary tract infections

Albert *et al.* (2004) reviewed 11 studies in which antibiotic treatments of recurrent urinary tract infections were compared to control groups. For more information, see Section 19.5. Here we only consider the second study, in which $X_2 = 8$ out of $n_2 = 21$ treated subjects continued to have infections while $X_1 = 17$ out of $n_1 = 19$ control subjects continued to have infections during the study period.

Letting p_1, p_2 be the probabilities (or risks) of further infection for the control and treated groups, we want the evidence for the alternative $p_2 < p_1$ to the null hypothesis $p_1 = p_2$. Letting $\Delta = p_1 - p_2$ we want the evidence for $\Delta > 0$.

Now $\tilde{p}_2 = 0.87500$, $\tilde{p}_1 = 0.38636$ and $\tilde{\Delta} = 0.48864$. Also $N = 40$, $q = n_2/N = 21/40 = 0.525$, so $\tilde{p} = 0.6429$ and $\tilde{\zeta} = 0.9206$.

This leads to $\tilde{\rho} = \tilde{\Delta}/\sqrt{\tilde{\zeta}} = 0.509$ and evidence $T = 3.38$ for $\Delta > 0$. Thus in this study there is moderate evidence for the antibiotic treatments being effective in reducing the risk of recurrent urinary tract infections.

A 95 % confidence interval for ρ is $[L, U] = [0.222, 0.748]$, for δ the interval is $[0.228, 1.125]$, for Δ it is $[0.213, 0.717]$ and for RR it is $[1.40, 3.69]$.

7.2.2 Diuretics in pregnancy and risk of pre-eclamsia

Collins *et al.* (1985) studied the benefit of taking diuretics during pregnancy on the risk of pre-eclamsia. The data were obtained in nine clinical trials. These data were also studied in Hardy and Thompson (1996) and in Biggerstaff and Tweedie (1997). The previous analyses concentrated on the odds ratio of developing pre-eclamsia. Here we consider for simplicity the differences in absolute risk $\Delta_k = p_1 - p_2$ for n_{2k} patients and n_{1k} controls, $k = 1, \cdots, 9$. The sample sizes are $N_k = n_{1k} + n_{2k}$. To calculate the evidence in each study we need the correlation effect sizes $\rho_k = \Delta_k/\sqrt{\zeta_k}$, where

Table 7.1 Results of nine independent randomized clinical trials of effect of diuretics on pre-eclamsia. For each study the number x_{1k} out of n_{1k} control subjects who had developed pre-eclamsia is listed, as well as the number x_{2k} of n_{2k} subjects treated with diuretics. See text for details regarding results.

k	x_{1k}	n_{1k}	x_{2k}	n_{2k}	$\tilde{\Delta}_k$	$\tilde{\zeta}_k$	$\tilde{\rho}_k$	L_k	U_k	$\hat{\kappa}_k$	T_k
1	14	136	14	131	−0.004	0.385	−0.006	−0.126	0.113	−0.006	−0.106
2	17	134	21	385	0.074	0.513	0.103	0.017	0.188	0.103	2.355
3	24	48	14	57	0.250	0.955	0.256	0.067	0.435	0.259	2.651
4	18	40	6	38	0.285	0.849	0.309	0.092	0.511	0.314	2.773
5	35	760	12	1011	0.034	0.126	0.097	0.050	0.143	0.097	4.070
6	175	1336	138	1370	0.030	0.411	0.047	0.009	0.085	0.047	2.454
7	20	524	15	506	0.008	0.134	0.023	−0.038	0.084	0.023	0.743
8	2	103	6	108	−0.036	0.159	−0.089	−0.222	0.045	−0.089	−1.299
9	40	102	65	153	−0.032	1.005	−0.032	−0.154	0.091	−0.032	−0.512

ζ_k are study-specific parameters. The evidence for a positive effect T_k was calculated from Equation (7.1). The data, estimated parameters $\hat{\zeta}_k$, correlation effect sizes $\hat{\rho}$ and 95 % confidence intervals for ρ, estimated transformed effects $\hat{\kappa}_k$ and evidence T_k from each trial are given in Table 7.1. The evidence is weak to moderate in studies 2–6, but there is no evidence of any benefit in the rest of the studies. For study 5 the difference in risk is rather small, $\Delta = 0.034$, but the evidence is the highest due to large sample sizes. The confidence interval for Δ is $[0.018, 0.051]$, the confidence interval for RR is $[1.82, 18.04]$ and the sample RR is 3.78. For study 4 the difference in risk is much larger, $\Delta = 0.285$, but the evidence is lower, this is a small study. The confidence interval for Δ is rather wide, $[0.085, 0.470]$, and the confidence interval for RR is $[1.32, 7.17]$, while the sample RR $= 2.71$.

8

Evaluating Poisson rates

Poisson processes are frequently used for modeling the number of rare events in time or space, such as numbers of equipment failures in a month, of traffic accidents at an intersection in a year, of mutations in a given segment of DNA, of cells growing in a culture (see Section 17.3.1), of soldiers killed by horse-kicks in each corps in the Prussian cavalry during 20 years (see Section 8.2.1), and a multitude of other phenomena. These numbers are called *counts data*, and one wants to use them to make inferences regarding the *rate of occurrence* of the rare events. The presentation below will be for processes in time.

8.1 Methodology

Data and model

- Given the observed number s_t of occurrences (the count) of the rare event during a time interval of known length t.

- For a Poisson process, the number S_t of events in any time interval of length t has a Poisson distribution with parameter μt, where $\mu > 0$ is the unknown rate of events per unit time.

- The Poisson(np) distribution also arises as an approximation to the binomial (n, p) distribution with large n and small probability p; this approximation is discussed in Section 18.4. One can think of n as a 'time' parameter and p, the probability of the rare event in any Bernoulli trial, as the rate per trial that the rare events occur.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

- One wants to make inferences regarding μ, which is estimated by the number of rare events in an observation period of length t, divided by the length of the interval $\hat{\mu} = S_t/t$. The mean and variance of the Poisson distribution equal its parameter, in this case μt, so $\hat{\mu}$ is an unbiased estimator of μ with variance μ/t. For large t, $\hat{\mu}$ is approximately normal with these parameters. The notation for the discrete time case in which S_n is the count of rare events in n trials (and is modeled by the Poisson(np) distribution) is $\hat{p} = S_n/n$.

Questions

- What is the evidence against the null hypothesis $H_0 : \mu = \mu_0$ and for the alternative $\mu > \mu_0$? Or, equivalently, for a positive effect $\Delta = \mu - \mu_0$?
- What is a confidence interval for Δ or for μ?

Test statistic and distribution

- The test statistic S_t has the Poisson(μt) distribution; large values of S_t favor $\Delta > 0$.
- A large-sample test statistic is

$$Z_t = \sqrt{\frac{t}{\mu_0}}\,(\hat{\mu} - \mu_0). \tag{8.1}$$

  Larger values of this test statistic favor the alternative $\Delta > 0$ over the null $\Delta = 0$. As the length of the observation period t increases without bound, the distribution of Z_t can be approximated by a normal distribution with mean $\sqrt{t}\,\delta$ and variance μ/μ_0, where $\delta = (\mu - \mu_0)/\sqrt{\mu_0}$ is the standardized effect. Under the null hypothesis, $\delta = 0$ and the approximating distribution of Z_t is standard normal.

Transformation to evidence

- The objective is to transform the test statistic onto the unit normal calibration scale; this is achieved by

$$T_t = 2\left\{\sqrt{S_t + 3/8} - \sqrt{\mu_0 t + 3/8}\right\}. \tag{8.2}$$

  This definition of evidence is based on the variance stabilizing transformation (Anscombe 1948) for the Poisson model (see Section 17.3.3).

- For $\mu t \geq 5$ the variance of T_t is stabilized near 1 (see Section 17.3.2).
- For each value of $\mu \geq \mu_0$ the Key is given by

$$\mathcal{K}(\mu|\mu_0) = 2(\sqrt{\mu} - \sqrt{\mu_0}). \tag{8.3}$$

- For large t the distribution of T_t is approximately normal with mean $\mathrm{E}[T_t] \doteq \sqrt{t}\,\mathcal{K}(\mu|\mu_0) = 2\sqrt{t}\,(\sqrt{\mu} - \sqrt{\mu_0})$ and unit variance. Under the null hypothesis, $\mu = \mu_0$ and $T_t \sim N(0, 1)$.

Interpretation

- The Key $\mathcal{K}(\mu|\mu_0)$ given by Equation (8.3) measures the distance from μ to μ_0. It is a simple function of the rate μ. It can be rewritten $\mathcal{K}(\mu|\mu_0) = 2(\sqrt{\Delta + \mu_0} - \sqrt{\mu_0})$ to show it is transformed effect; that is, a function of the effect Δ.
- It is also possible to rewrite the Key $\mathcal{K}(\mu|\mu_0)$ as a function of the standardized effect δ, as is usually done in this book. For the Poisson model this only complicates the Key without adding any insight.
- An inverse transformation to find μ from the transformed effect is

$$\mathcal{K}^{-1}(y|\mu_0) = [\,\max\{\,(y/2 + \sqrt{\mu_0}\,), 0\,\}\,]^2. \tag{8.4}$$

  Since the rate $\mu \geq 0$ the values of $\mathcal{K}^{-1}(y|\mu_0)$ are truncated at zero.

- The approximate power of the level-α test based on the evidence T_t for detecting an alternative $\mu_1 > \mu_0$ is

$$1 - \beta(\mu_1|\mu_0) = \Phi(\sqrt{t}\,\mathcal{K}(\mu_1|\mu_0) - z_{1-\alpha}).$$

  This formula can be rewritten to give the expected evidence in terms of level and power $\tau = z_{1-\alpha} + z_{1-\beta(\mu_1|\mu_0)}$.

Choosing the time *t* required to achieve a desired amount of evidence

- To obtain expected evidence $\tau_1 = \sqrt{t}\,\mathcal{K}(\mu_1|\mu_0)$ for alternative μ_1, the required observation time is $t = \{\tau_1/\mathcal{K}(\mu_1|\mu_0)\}^2$.

Confidence intervals

- Let $c = z_{0.975} = 1.96$. A nominal 95 % confidence interval for $\mathcal{K}(\mu|\mu_0)$ is given by

$$[L, U] = t^{-1/2}[T_n - c\,, T_n + c\,]. \tag{8.5}$$

- A nominal 95 % confidence interval for μ is given by

$$[L_\mu, U_\mu] = t^{-1/2}\left[\mathcal{K}^{-1}(T_t - c|\mu_0)\,, \mathcal{K}^{-1}(T_t + c|\mu_0)\right], \tag{8.6}$$

  where the function $\mathcal{K}^{-1}(y|\mu_0)$ is given by Equation (8.4). This interval is not symmetric around μ; its lower limit is non-negative due to the restrictions imposed on the inverse key function; the left limit $L_\mu \geq \mu_0$ if and only if $T_t \geq c$. The coverage of this interval is good when $t\mu \geq 5$, the range usually encountered in applications.

8.2 Example

8.2.1 Deaths by horse-kicks

This famous example of numbers of soldiers killed per year by horse-kicks in each corps in the Prussian cavalry belongs to L.J. Bortkiewicz (1868–1931) and has been used to illustrate the application of the Poisson distribution in numerous textbooks. These data are taken from Preece *et al.* (1988), who reproduce the complete data on number of deaths each year for 20 years (1875 to 1894) for each of 14 corps. After omitting four anomalous corps, they fit a log–linear model to the data and show that the death rates are constant over time, but with different rates for different corps. The total S_{20} number of deaths during the 20 years for each corps are displayed in the second column of Table 8.1.

For the sake of illustration assume that the safety regulations in the Prussian cavalry required that the death rate from horse-kicks to be at most 0.5 per year in a corps. Then $\mu_0 = 0.5$ and to ascertain compliance for each corps one needs to test

Table 8.1 Bortkiewicz's data on number S_t of deaths by horse-kicks over 20 years for each of 10 corps. The large-sample test statistic Z_t is defined in Equation (8.1) and the evidence T_t is calculated with Equation (8.2). The p-values p_z and p_T are for the tests based on Z_t and T_t, respectively, using the normal distribution. L_μ and U_μ are the lower and upper limits of the 95 % confidence interval for the unknown rate μ. The length of the observation period is $t = 20$ for each individual corps, $t = 200$ for the combined 10 corps (the row marked 'Total') and $t = 180$ after omitting corps XIV (the last row).

Corps	S_t	$\hat{\mu}$	Z_t	p_Z	T_t	p_T	L_μ	U_μ
II	12	0.6	0.632	0.264	0.594	0.276	0.31	0.99
III	12	0.6	0.632	0.264	0.594	0.276	0.31	0.99
IV	8	0.4	−0.632	0.736	−0.654	0.743	0.00	0.73
V	11	0.55	0.316	0.376	0.303	0.381	0.00	0.92
VII	12	0.6	0.632	0.264	0.594	0.276	0.31	0.99
VIII	7	0.35	−0.949	0.829	−1.011	0.844	0.00	0.66
IX	13	0.65	0.949	0.171	0.872	0.192	0.34	1.05
X	15	0.75	1.581	0.057	1.400	0.081	0.42	1.17
XIV	24	1.2	4.427	0.000	3.432	0.000	0.76	1.72
XV	8	0.4	−0.632	0.736	−0.654	0.743	0.00	0.73
Total	122	0.61	2.200	0.014	2.087	0.018	0.51	0.72
Total w/o XIV	98	0.54	0.843	0.200	0.824	0.205	0.44	0.66

the null $\mu = 0.5$ against the alternative $\mu > 0.5$. The results are given in columns 3 to 9 of Table 8.1.

Several corps have estimated death rates $\hat{\mu}$ above $\mu_0 = 0.5$, but the evidence for noncompliance is negligible except for corps X, where it is very weak ($T_{20} = 1.4$), and for corps XIV, where it is moderate ($T_{20} = 3.43$).

But worryingly the overall death rate for all 10 corps (in the row marked total) is too high; there the evidence $T_{200} = 2.09$ is weak, but certainly not negligible. The computations are repeated after omitting the data for corps XIV and the results listed in the last row of the table; this time there is negligible evidence for overly high death rates.

The p-values from the Z-test (p_Z) and those based on the evidence (p_T) are provided in Table 8.1 to compare the quality of the usual normal approximation to the Poisson distribution with that of the *vst*-based approximation. The evidence-based p-values are somewhat higher. This difference is noticeable for corps X, where $p_Z = 0.057$ (rather close to being significant at level 0.05) whereas $p_T = 0.081$. Now S_{20} has the null distribution Poisson(10), so the exact p-value to three decimal places is $P(S_{20} \geq 15) = 0.083$.

Similarly, for the total number of deaths, one finds $p_Z = 0.014$, $p_T = 0.018$ and the exact p-value to three decimal places is 0.018. The exact p-values are much closer to p_T than to p_Z because the variance stabilization provides a superior approximation to the Poisson distribution compared to the standard normal approximation. The coverage of the evidence-based confidence intervals for μ is also much more reliable.

9

Comparing Poisson rates

For two-sample data modelled by Poisson distributions with unknown rates μ_1, μ_2 one is often interested in comparing these two parameters. One may do this with either the difference $\Delta = \mu_2 - \mu_1$ or the ratio $\rho = \mu_2/\mu_1$. Testing for the one-sided alternative $\Delta > 0$ to the null $\Delta = 0$ is equivalent to testing for alternative $\rho > 1$ to the null $\rho = 1$, but the statistical methods are different for these two parameterizations of the problem.

For the difference Δ, one first finds transformed parameter estimates of each rate; this leads to evidence for the alternative hypothesis $\Delta > 0$ whose expectation depends on a standardized effect δ, from which confidence intervals for δ can be derived. This is called the *unconditional approach* in what follows.

For the ratio ρ the conditional distribution of the second one sample total, given the two-sample total, is the basis for inference. It leads to evidence for the alternative hypothesis whose expectation depends on ρ, from which confidence intervals for ρ can be derived. This is called the *conditional approach*.

Recall that the Poisson(np) distribution is often used as an approximation to a binomial(n, p) distribution with large n and small probability p. The parameter p is the rate of occurrence of the rare event per Bernoulli trial, and is often called a risk (of infection, disease, death) in the medical literature. When this approximation is applicable to each of two samples with respective rates p_1, p_2, the difference of rates $\Delta = p_2 - p_1$ is called the risk difference and the ratio of rates $\rho = p_2/p_1$ is called the risk ratio or relative risk. Confidence intervals for the risk difference under the Binomial model are found in Chapter 7.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte
© 2008 John Wiley & Sons, Ltd

9.1 Methodology

9.1.1 Unconditional evidence

Data and model

- Rare events are observed to occur over time under two different sets of conditions. The observed numbers x_{t_1}, x_{t_2} of these rare events in time intervals of respective lengths t_1, t_2 are recorded.

- The measurements are modelled by independent Poisson processes with respective rate parameters μ_1, μ_2. Thus for the first process the number X_{t_1} of rare events has a Poisson distribution with parameter $\mu_1 t_1$, and similarly for the second process Y_{t_2} ~Poisson $(\mu_2 t_2)$.

Questions

- What is the evidence for a positive effect $\Delta = \mu_2 - \mu_1$; or, equivalently, for a positive standardized effect $\delta > 0$ defined by Equation (9.2)?

- What is a confidence interval for Δ or for δ?

Test statistic and transformation to evidence

- A natural test statistic is $\hat{\Delta} = \hat{\mu}_2 - \hat{\mu}_1 = Y_{t_2}/t_2 - X_{t_1}/t_1$, but its distribution is complicated and the standardized version $(\hat{\Delta} - 0)/\text{SE}[\hat{\Delta}]$ converges to the standard normal distribution much slower than the evidence statistic T_t defined below and first introduced by Huffman (1984).

- As in the previous chapter, one can stabilize the variances of X_{t_1}, Y_{t_2} to one by applying the Anscombe transformation to obtain the statistics $S_X{=}2\sqrt{X_{t_1}{+}3/8}$ and $S_Y{=}2\sqrt{Y_{t_2}{+}3/8}$. Then, letting $t = t_1 + t_2$, and $q = t_2/(t_1 + t_2)$, the evidence for $\Delta > 0$ is defined by

$$T_t = \sqrt{1-q}\, S_Y - \sqrt{q}\, S_X . \tag{9.1}$$

- For large t_1, t_2 the distribution of T_t is approximately normal with mean $\text{E}[T] \doteq \sqrt{t}\,\delta$ and variance 1, where the *standardized effect* δ is defined by

$$\delta = 2\sqrt{q(1-q)}\,(\sqrt{\mu_2} - \sqrt{\mu_1}). \tag{9.2}$$

  Under the null hypothesis, $\delta = 0$ and for large t_1, t_2, the statistic T_t is approximately standard normal.

- The Huffman (1984) statistic defined by (9.1) is a *vst* for this two-sample Poisson problem and because $\text{E}[T_t] \doteq \sqrt{t}\,\delta$, the Key Inferential Function is especially simple, $\mathcal{K}(\delta) = \delta$.

- For $t_1\mu_1 \geq 5$ and $t_2\mu_2 \geq 5$ the variances of S_X and S_Y are reliably stabilized near 1 (see Section 17.3.2); therefore the same is true for the test statistic T_t, which gives the evidence for $\delta > 0$.

Interpretation

- The standardized effect δ given by Equation (9.2) is a comparatively simple function of the individual parameters μ_1 and μ_2. But it cannot be rewritten as a function of the difference Δ alone or the ratio ρ alone. Thus the statistic T_t is a good measure of evidence, but it is not especially useful in itself for finding confidence intervals for the risk difference or risk ratio.
- The approximate power of the level-α test based on T_t for detecting an alternative $\delta_1 > 0$ is

$$1 - \beta(\delta_1) = \Phi(\sqrt{t}\,\delta_1 - z_{1-\alpha}).$$

  This formula can be rewritten to give the expected evidence in terms of level and power $\tau = z_{1-\alpha} + z_{1-\beta(\delta_1)}$.

Choosing the sample size to achieve a desired amount of evidence

- To obtain expected evidence $\tau = \sqrt{t}\,\delta_1$ for a standardized effect δ_1 of scientific interest when this effect actually exists, the total observation time $t = t_1 + t_2$ must satisfy $t \geq \{\tau/\delta_1\}^2$. Individual observation times are found from $t_1 = (1-q)t$ and $t_2 = qt$. It is clear from Equation (9.1) that the maximum evidence for a given t is obtained for $q = 0.5$.

Confidence intervals

- Let $c = z_{0.975} = 1.96$. A nominal 95 % confidence interval for δ is given by

$$[L, U] = t^{-1/2}[T - c\,, T + c\,]. \tag{9.3}$$

  The coverage of this interval is good when $\mu_1 t_1 \geq 5$ and $\mu_2 t_2 \geq 5$. For fixed q the range of good coverage improves with t.

9.1.2 Conditional evidence

Data and model

- The data and model are exactly as in the previous section, but now the ratio of rates $\rho = \mu_2/\mu_1$ is of interest. Inference is conditional on the observed total $X_{t_1} + Y_{t_2} = m$ which is assumed fixed.

Questions

- What is the evidence for alternative $\rho > 1$ to the null $\rho = 1$?
- What is a confidence interval for ρ ?

Test statistic and distribution

- The conditional distribution of Y_{t_2} given $X_{t_1} + Y_{t_2} = m$ is binomial with parameters m and $p = t_2\mu_2/(t_1\mu_1 + t_2\mu_2) = (1 + (q^{-1} - 1)\rho^{-1})^{-1}$, where $q = t_2/(t_1 + t_2)$ and $\rho = \mu_2/\mu_1$ (see Lehmann (1986), pp.140–142). Let $Y_{t_2|m} \sim \text{binomial}(m, p)$ be a random variable with this conditional distribution; it is the test statistic for inference on p.
- The parameter p is a monotonically increasing function of ρ, with inverse function $\rho = (q^{-1} - 1)(p^{-1} - 1)^{-1}$. The hypotheses $\rho = 1$ versus $\rho > 1$ are equivalent to $p = q$ versus $p > q$. Large values of $Y_{t_2|m}$ favor the alternative $p > q$, so the traditional conditional test is carried out for an observed $Y_{t_2|m} = y$ by computing and evaluating the p-value $P(Y_{t_2|m} \geq y \mid p = q)$. For large t_1, t_2 the statistic $Y_{t_2|m}$ is approximately normal with mean mp and variance $mp(1 - p)$.

Transformation to evidence

- The conditional evidence T_{cond} for the alternative $p > q$, and hence for $\rho > 1$ is obtained by applying the *vst* for binomial distributed variables described in Section 18.1 to S_Y:

$$T_{\text{cond}} = 2\sqrt{m}\,\{\arcsin(\sqrt{\tilde{p}}\,) - \arcsin(\sqrt{q}\,)\}, \tag{9.4}$$

where $\tilde{p} = (Y_{t_2|m} + 0.375)/(m + 0.75)$. Another expression is obtained via the trigonometric identity $\arcsin(\sqrt{x}\,) + \arcsin(1 - 2x) = \pi/2$:

$$T_{\text{cond}} = 2\sqrt{m}\,\{\arcsin(1 - 2q) - \arcsin(1 - 2\tilde{p})\}. \tag{9.5}$$

- It follows from the properties of the *vst* transformed test statistic $Y_{t_2|m}$ that T_{cond} is approximately normal with variance 1 and mean $\mathrm{E}[T] \doteq \sqrt{m}\,\mathcal{K}(p)$, where the Key is given by

$$\mathcal{K}(p) = \arcsin(1 - 2q) - \arcsin(1 - 2p). \tag{9.6}$$

Interpretation

- Recall that the binomial parameter $p = p(\rho)$ is monotonically increasing in the risk ratio ρ. Therefore one can interpret T_{cond} as conditional evidence for the alternative $\rho > 1$ to the null $\rho = 1$.

- The approximate power of the level-α test for detecting an alternative $\rho_1 > 1$ can be found exactly as in Section 9.1.1.

Confidence intervals

- Let $c = z_{0.975} = 1.96$. A nominal 95 % confidence interval for $\kappa = \mathcal{K}(p)$ is given by

$$[L, U] = m^{-1/2}\left[T_{\text{cond}} - c\,,\, T_{\text{cond}} + c\right]. \tag{9.7}$$

- Let $C = \arcsin(1 - 2q)$ and write the inverse function to $\kappa = \mathcal{K}(p(\rho))$

$$h(\kappa) = \frac{(1-q)(1-\sin(C-\kappa))}{q(1+\sin(C-\kappa))}. \tag{9.8}$$

- A nominal 95 % confidence interval for ρ is $[h(L), h(U)]$.
- The coverage of the above intervals is good for $mp(1-p) \geq 5$.

Application to Bernoulli trials data

- In the preamble to this chapter it was noted that given a large number n_1 of Bernoulli trials indicating the occurrence of disease or no disease, say, and each trial resulting in disease with small risk p_1, the Poisson($n_1 p_1$) approximation could be a model for the number X_{n_1} in the sample that have the disease. Let Y_{n_2} be the number in an independent sample of large size n_2, small risk p_2, so that Y_{n_2} ~Poisson($n_2 p_2$). Then it is clear that the above conditional methods (with $t_i = n_i$, $\mu_i = p_i$) apply to the relative risk $\rho = p_2/p_1$. The unconditional methods provide evidence for $\Delta = p_2 - p_1 > 0$ and confidence intervals for $\sqrt{p_2} - \sqrt{p_1}$. Confidence intervals for Δ based on the Binomial model are found in Chapter 7.

9.2 Example

9.2.1 Vaccination for the prevention of tuberculosis

The data in Table 9.1 are reproduced from Sutton *et al.* (2000). The data resulted from 13 randomized clinical trials (RCTs), each comparing a group vaccinated by Bacillus Calmette-Guerin (BCG) vaccine for the prevention of tuberculosis against a nonvaccinated group, and originally reported by Colditz *et al.* (1994). It was suspected that the distance from the equator affected the efficacy of the vaccine, and therefore this covariate is investigated by means of meta-regression in Chapter 14.

The sample sizes in the RCTs are large and the risks of tuberculosis are relatively small, so we model the data using the Poisson distribution. Here unconditional and conditional evidence for each trial is calculated, and the two are compared. Also calculated are confidence intervals for the relative risk ρ of tuberculosis in the nonvaccinated group. All results are given in Table 9.2.

Table 9.1 Data from clinical trials of BCG vaccine efficacy, reproduced in a modified form from Colditz *et al.* (1994).

		Vaccinated		Not vaccinated	
Trial	Latitude	Disease	No disease	Disease	No disease
1	44	4	119	11	128
2	55	6	300	29	274
3	42	3	228	11	209
4	52	62	13 536	248	12 619
5	13	33	5036	47	5761
6	44	180	1361	372	1079
7	19	8	2537	10	619
8	13	505	87 886	499	87 892
9	−27	29	7470	45	7232
10	42	17	1699	65	1600
11	18	186	50 448	141	27 197
12	33	5	2493	3	2338
13	33	27	16 886	29	17 825

The trials in Table 9.1 vary considerably in both sample size and number of diseased. The total number of cases m varies from 8 (trial 12) to 1004 (trial 8), and the total number of subjects $N = n_1 + n_2$ varies from 262 (trial 1) to 176 782 (trial 8). Unconditional and conditional evidence for a relative risk $\rho > 1$ are listed in the second and third columns of Table 9.2 and they nearly coincide. This is possibly

Table 9.2 Unconditional and conditional evidence, relative risk and its confidence interval for the data from clinical trials of BCG vaccine efficacy.

	Unconditional	Conditional		95 % CI for RR	
Trial	evidence	evidence	RR	Lower	Upper
1	1.57	1.55	2.43	0.95	6.95
2	4.12	4.16	4.88	2.40	10.79
3	2.24	2.26	3.85	1.37	12.80
4	11.58	11.78	4.23	3.36	5.36
5	0.95	0.95	1.24	0.86	1.81
6	8.99	9.04	2.19	1.89	2.55
7	3.19	3.20	5.06	2.29	11.36
8	−0.19	−0.19	0.99	0.89	1.10
9	1.99	1.98	1.60	1.08	2.38
10	5.67	5.74	3.94	2.54	6.27
11	3.00	3.00	1.40	1.17	1.69
12	−0.59	−0.56	0.64	0.17	2.19
13	0.06	0.06	1.02	0.65	1.59

due to reasonably large values of N. Usually, conditional tests are considered to be less powerful than unconditional, but we do not see this here. Evidence varies from negligible (trials 1, 5, 8, 12, 13) to strong (trials 4, 6, 10). The null hypothesis of equal risks is rejected by traditional tests whenever the evidence is above 1.65.

The 95 % confidence intervals for the relative risk ρ in Table 9.2 are based on conditional inference, and are calculated from Equation (9.8). Note that the strength of evidence and RR are not directly related: evidence also very much depends on the total number of cases. Evidence in trial 6 is higher than evidence in trial 10, even though the relative risk is only half as big. This is due to $m = 552$ versus $m = 82$ in the respective trials.

10

Goodness-of-fit testing

Given observations $x_1, \ldots, x_N$ from an unknown distribution, it is often desired to think of them as a sample from a population which has shape similar to some standard distribution F, where F may or may not depend on unknown parameters. This problem is called goodness-of-fit testing, meaning that the test is required to see how well the distribution F fits the data. Only one test, the classic Pearson's chi-square test, is discussed in this chapter. We treat the case of fully known F foremost and only briefly comment on the case of estimated parameters.

We do not recommend the indiscriminate use of the chi-square test in goodness-of-fit problems. Its main disadvantage is the arbitrariness of the choice of the number of intervals K (see below). There are better goodness-of-fit tests, such as the Anderson–Darling test or Shapiro–Wilks test for the normal distribution (see D'Agostino and Stephens, 1986). Rather, this chapter is included to demonstrate the breadth of possible applications of our approach to evidence.

10.1 Methodology

Data and model

- Let $X_1, \ldots, X_N$ be independent observations, each with the same unknown distribution F, and partition the domain of F into K intervals, with p_k equalling the probability under F of an observation falling within the kth interval. Denote by O_k the observed number of observations falling in the kth interval. The expected number of observations falling in the kth interval is calculated as $E_k = Np_k$.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

Questions

- Are the O_k's close enough to the respective E_k's so that the hypothesized model F can be adopted? Or is there sufficient evidence to reject the model?
- Are the O_k's too close to the respective E_k's? That is, is there another explanation for the very close agreement between model and data?

Test statistic and distribution

- Pearson's chi-squared statistic $\mathcal{C}$ is defined by

$$\mathcal{C} = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}. \tag{10.1}$$

- Pearson's chi-squared statistic can be rewritten to bring more insight into its properties. Under the true distribution $F^{(N)}$ the probabilities $p_k^{(N)}$ of an observation falling within the kth interval are estimated by $\hat{p}_k^{(N)} = O_k/N$. The observed numbers $O_k = N\hat{p}_k^{(N)}$ are compared to the expected numbers $E_k = Np_k$. Then the statistic is

$$\mathcal{C} = N \sum_{k=1}^{K} \frac{\left(\hat{p}_k^{(N)} - p_k\right)^2}{p_k}.$$

- Under certain regularity conditions (see Chapter 3 of Greenwood and Nikulin (1996)), for large N the null distribution of $\mathcal{C}$ is approximately χ^2_{K-1}; and Pearson's test rejects the hypothesized model F at level α when $\mathcal{C} \geq \chi^2_{K-1,0.95}$.
- Under alternatives (see p. 23 of Greenwood and Nikulin (1996)), the Pearson statistic $\mathcal{C}$ of Equation (10.1) has for large N approximately a $\chi^2_{K-1}(\lambda_N)$ distribution, where

$$\lambda_N = N \sum_{k=1}^{K} \frac{\left(p_k - p_k^{(N)}\right)^2}{p_k}. \tag{10.2}$$

Transformation to evidence

- Evidence in the noncentral chi-squared statistic is found in Section 22.2. The large-sample evidence in $\mathcal{C}$ for such alternatives $\lambda > 0$ to $\lambda = 0$ is obtained from Equation (22.1), namely

$$T_{K-1} = h_{K-1}(\mathcal{C}). \tag{10.3}$$

- Thus rather than carrying out a traditional test and computing power for various alternatives, one can evaluate and interpret the evidence T_{K-1} which has an approximate normal distribution, with standard error 1 in estimating its mean defined in Equation (22.2). In view of remark 2 following Definition 22.3, for large N a simple-to-remember version of this Key is $\mathcal{K}(\theta) = \theta^{1/2}$, where $\theta = \lambda/N$, and the expected evidence $\mathrm{E}[T_{K-1}] \doteq \sqrt{N}\,\mathcal{K}(\theta) = \lambda^{1/2}$.

Interpretation

- As usual, we recommend that T_{K-1} equal to 1.645, 3.3 and 5, each with standard error 1, be roughly interpreted as weak, moderate and large evidence against the model. A weak, moderate and large negative evidence is interpreted as an evidence of a 'too good to be true' model fit. If T_{K-1} were -3.3, say, this result would be interpreted as moderate evidence for the existence of particular reasons for an extraordinarily good fit between model and data.
- Greenwood and Nikulin (1996, pp. 27–28) remark that point hypotheses such as $\lambda_0 = 0$ never really hold, and that one should really be testing $\lambda \leq \lambda_0$ against $\lambda > \lambda_0$, for some suitable choice of λ_0. This amounts to relaxing the model assumption, and making it harder to reject the model(s) represented by $p_1, \ldots, p_K$. They state that in many practical applications one can choose $\lambda_0 = (1/NK^2)\sum_k(1/p_k)$; and, in particular for equiprobable intervals $\lambda_0 = 1/N$. Were this suggestion to be adopted, the appropriate measure of evidence from Definition 22.6 would be given by $T_{K-1}(\lambda_0) = T_{K-1} - \sqrt{\lambda_0} = T_{K-1} - 1/\sqrt{N}$. In practice this theoretical nicety is not likely to make much difference to the evidence obtained.
- When the parameters of the distribution F need to be estimated from the data, and these estimates are then used to calculate the expected values E_k in the Pearson statistic, it is commonly stated that $\mathcal{C} \sim \chi^2_{K-1-r}$, where r is the number of estimated parameters. In fact this is true only when the parameters are estimated from the cell counts O_k and not from the original observations x_k. In the latter case, the distribution will lie somewhere between a chi-square distribution with $K - r - 1$ and $K - 1$ degrees of freedom (see Chernoff and Lehmann 1954). Usually this results in the inflated level of the goodness-of-fit test above the nominal level. Luckily, the difference is not too large when the number of intervals K is large. Watson (1957) recommends $K > 10$ for a normal case with the mean and variance estimated by their sample counterparts.

Choosing the sample size

- Sample size calculations for obtaining an expected evidence for a fixed number of intervals K and alternative $\theta_1 = \lambda/N$ are discussed in detail in Section 22.4.1. Alternatively, sample size calculations required to obtain a desired power are given in Section 22.4.2. Such calculations are of not much value here, because the choice of K is arbitrary and greatly affects the results.

10.2 Example

10.2.1 Bellbirds arriving to feed nestlings

Table 10.1 gives the collective arrival times of bellbirds to a nest containing two 10-day-old nestlings. The mother bird is the only female amongst the dozen bellbirds arriving during a 90-minute period. The father of the nestlings is quarantined by the zoologists, so the males can be considered potential suitors, trying to impress the female by their paternal efforts in feeding the nestlings.

It would be easier to test this hypothesis if one could model the arrival times as a Poisson process in time. For such a process, given that the number of arrivals during the observation period is N, the arrival times have the same joint distribution as if they were drawn independently from the uniform model on this interval (see Karlin and Taylor (1975), p. 126). This result is the basis for a goodness-of-fit test for a Poisson process. Are the N arrival points distributed 'at random' within the interval? That is, are they consistent with the uniform model or are they too regularly spaced or too clumped together to be consistent with a Poisson(λ) process?

The bellbird arrival data do appear more or less uniformly distributed over the 90 minutes, with small gaps at 40 and 85 minutes. The gap at 40 minutes can be partially explained by the arrival of a much larger bird, a currawong.

10.2.1.1 Testing for 'randomness' of arrivals

The method is to partition the observation interval into K intervals of equal length, each having probability $1/K$ under the uniform model, and test for 'randomness' by comparing the expected number in each interval $E_k = N/K$ with the observed number O_k via the statistic $\mathcal{C} = \sum_k (O_k - E_k)^2/E_k$. In this case both large values

Table 10.1 Arrival times of bellbirds to a nest during a 90-minute observation period, after conversion from minutes and seconds to decimal notation. The data were kindly supplied by Dr Michael Clarke of the School of Zoology, La Trobe University.

0.000	0.250	2.250	2.583	2.750	3.200	3.500
3.750	4.583	5.583	5.666	6.500	6.666	7.000
7.750	9.166	9.750	10.000	11.083	15.666	15.833
16.500	17.833	18.083	18.916	19.416	21.833	22.666
22.750	24.133	24.750	26.166	26.250	26.500	28.500
29.916	30.333	30.583	32.333	33.166	33.500	34.333
35.166	35.666	36.250	36.500	38.833	39.000	43.583
44.500	49.000	50.833	50.916	51.000	51.250	51.833
52.166	53.250	54.333	55.000	55.333	56.916	57.333
57.666	57.916	58.666	58.750	59.916	60.716	62.916
65.500	66.250	66.833	67.250	67.916	69.916	71.333
72.250	73.166	76.583	77.583	78.583	78.750	79.750
81.666	82.483	88.916				

of $\mathcal{C}$ and small values of $\mathcal{C}$ raise doubt about the hypothesis of a Poisson process. Large values suggest finding another model which explains the agglomeration of observations, while small values suggest finding a model which explains why the arrivals are so regular.

An analogous test for complete spatial randomness is often carried out when observations are points in the plane, such as locations of trees in a forest, or of cells growing in culture in a dish. The observation region is partitioned into a grid of K equal-area regions, and the number of points falling within each region are the observations O_k. See Diggle (1983) for examples.

Returning to the bellbird data, there are $N = 87$ observations in the 90-minute period. If one uses $K = 6$ intervals of length 15 minutes each, the expected number in each interval is $E_k = 87/6 = 14.5$. The observed numbers O_k are 19, 17, 14, 18, 11 and 8, respectively, leading to a chi-squared statistic $\mathcal{C} = 6.45$ and $T_5 = h_{87}(6.45) = 0.59$. This result has standard error 1, so can be considered neutral, discrediting neither the uniform distribution of the arrivals, nor the hypothesis of randomness of arrivals. If $\mathcal{C}$ had been 11.9, then $T_5(\mathcal{C}) = +1.645$ and we would say there is weak evidence against the model in favor of a model which allows for more agglomeration of points than a uniform model. The traditional Pearson chi-squared test rejects at level 0.05 for $\mathcal{C} = 11.07$.

If the result had turned out near $\mathcal{C} = 0.98$ with $T_{87} = h_{87}(0.98) = -1.645$, or if $\mathcal{C} = 0.09$ with $T_{87} = h_{87}(0.09) = -3.3$, then we would say there is weak or moderate evidence for the nonrandomness, and look for a specific reason why there is so much evidence: possibly the birds avoid each other, so the gaps between arrivals are more regular than would be expected from random arrivals.

11

Evidence for heterogeneity of effects and transformed effects

Given K studies, it is customary in the meta-analytic literature to carry out a chi-squared test of the hypothesis of homogeneity of effects using Cochran's Q statistic, introduced by Cochran (1937, 1954). If the test fails to reject it is then assumed the effects are equal, and an estimate of the common effect can be obtained; if it does reject then an alternative model which allows for different effects is assumed. In this chapter we measure the evidence for the alternative of unequal effects, and also evidence for unequal transformed effects. In either case, the evidence for heterogeneity is different for fixed and random effects models, so will be presented in separate sections below. The theory is given in Chapter 24.

11.1 Methodology

11.1.1 Fixed effects

Data and Model

- Given K studies of sizes n_k measuring potentially different effects μ_k, for $k = 1, \dots, K$.
- The estimated effects $\hat{\mu}_k$ for the respective studies are mutually independent and approximately normal with variances w_k^{-1}.
- The inverse variances w_k are used as weights for the effects μ_k, and their estimates are denoted by $\hat{w}_k$.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

- Denote the weighted mean effect by $\bar{\mu}_w = \sum w_k \mu_k / \sum w_k$ and its estimate $\hat{\bar{\mu}}_{\hat{w}} = \sum \hat{w}_k \hat{\mu}_k / \sum \hat{w}_k$. For equal effects $\mu_k = \bar{\mu}_w$ for all k.
- Standardized effects are denoted by δ_k, and transformed (standardized) effects by $\kappa_k = \mathcal{K}(\delta_k)$, where $\mathcal{K}$ is the Key for the model, which is assumed to be the same for all studies. These transformed effects can be combined with weights n_k to obtain $\kappa = \sum n_k \kappa_k / N$, where $N = \sum n_k$ is the total sample size.
- Evidence in the kth study satisfies $T_k \sim N(\sqrt{n_k}\kappa_k, 1)$, to a good approximation.
- Transformed effects are estimated by $\hat{\kappa}_k = T_k / \sqrt{n_k}$; for each k the estimator $\hat{\kappa}_k$ is approximately normal with mean κ_k and variance n_k^{-1}. Their weighted mean is $\hat{\kappa} = \sum n_k \hat{\kappa}_k / N$.

Questions

- What is the evidence against the null hypothesis of homogeneity of effects $H_0 : \mu_k = \mu$ for all k and for the alternative of heterogeneity $H_1 : \mu_j \neq \mu_k$ for some $j \neq k$?
- Alternatively, what is the evidence against the null hypothesis of homogeneity of *transformed* effects $H_0^* : \kappa_k = \kappa$ for all k and for the alternative $H_1^* : \kappa_j \neq \kappa_k$ for some $j \neq k$?

Test statistic and distribution

- To test for the homogeneity of effects, Cochran's Q is defined by

$$Q = \sum_k \hat{w}_k (\hat{\mu}_k - \hat{\bar{\mu}}_{\hat{w}})^2. \tag{11.1}$$

  Larger values of the test statistic favor the alternative H_1 of the heterogeneity of effects over the null H_0 of homogeneity.

- For a fixed number of studies K and simultaneously growing sample sizes $n_k \to \infty$ (see Section 24.1.1 for details), the distribution of Q is approximately $\chi^2_{K-1}(\lambda)$ with the noncentrality parameter $\lambda = \sum w_k (\mu_k - \bar{\mu}_w)^2$. Under the null hypothesis H_0, $\lambda = 0$ and Q has the central χ^2_{K-1} distribution.
- To test for the homogeneity of *transformed* effects, the test statistic Q^* is defined by

$$Q^* = \sum_k n_k (\hat{\kappa}_k - \hat{\kappa})^2. \tag{11.2}$$

  Larger values of Q^* favor the alternative H_1^* of heterogeneity of transformed effects over the null H_0^* of homogeneity.

- For a fixed number of studies K and simultaneously growing sample sizes $n_k \to \infty$, the distribution of Q^* is approximately $\chi^2_{K-1}(\lambda^*)$ with the noncentrality parameter $\lambda^* = \sum n_k(\kappa_k - \kappa)^2$. Under the null hypothesis H_0^*, $\lambda^* = 0$ and Q^* has the central χ^2_{K-1} distribution.

Transformation to evidence

- The evidence for heterogeneity in Q is defined by
$$T_Q = \sqrt{Q \ - m/2} - \sqrt{m/2}, \tag{11.3}$$
where $m = \chi^2_{K-1,0.5}$ is the null median. This formula only applies for $Q \geq m$. For Q less than its null median it is defined by a symmetrization argument (see Definition 24.1 in Section 24.1.2 for details).

- For large individual study sample sizes totaling N, Cochran's Q is approximately distributed as $\chi^2_{K-1}(\lambda)$. Then T_Q is approximately $N(\tau_Q, 1)$, where for $\theta = \lambda/N$ the expected evidence is $\tau_Q \doteq \sqrt{N}\,\mathcal{K}(\theta)$, and $\mathcal{K} = \mathcal{K}_{K-1,N}$ is given by (22.2). This Key for the noncentral chi-squared model is very complicated, but for λ, N increasing without bound and λ/N approaching θ, the Key approaches $\mathcal{K}(\theta) = \sqrt{\theta}$. Under the null hypothesis H_0, $\tau_Q = 0$ and $T_Q \sim N(0, 1)$.

- The evidence in Q^* for heterogeneity of *transformed* effects is defined to be the *vst* in (22.1) applied to Q^*, and denoted T_{Q^*}. For large sample sizes $T_{Q^*} \sim N(\tau_{Q^*}, 1)$, where the mean evidence is given by $\tau_{Q^*} \doteq \sqrt{N}\,\mathcal{K}(\lambda^*/N)$. Under the null hypothesis H_0^*, $\tau_{Q^*} = 0$ and $T_{Q^*} \sim N(0, 1)$.

- In the normal model with equal sample sizes $n_k = n$ described in Section 24.2 the variance of T_Q is reliably stabilized near 1 for $n \geq 80$, and the variance of T_{Q^*} for $n \geq 20$.

Interpretation

- Hypotheses (H_0, H_1) and Cochran's test based on Q are concerned with heterogeneity of effects μ_k, whereas hypotheses (H_0^*, H_1^*) and the test based on Q^* are concerned with heterogeneity of *transformed* effects κ_k. In general, these two problems are different. It is possible to have homogeneous effects and heterogeneous transformed effects, or the other way round.

- The approximate power of the level-α T_Q-based test for detecting an alternative $\lambda > 0$ is
$$1 - \beta(\lambda) = \Phi(\sqrt{N}\,\mathcal{K}(\lambda/N) - z_{1-\alpha}).$$
This formula can be rewritten to give the expected evidence in terms of level and power: $\tau = z_{1-\alpha} + z_{1-\beta(\lambda)}$.

- The usual methodology can be used for sample size calculations, and confidence intervals for the noncentrality parameter λ derived. Neither seems to be of much practical interest.

11.1.2 Random effects

Data and model

- Continuing with the assumptions and notation of Section 11.1.1, further assume $\mu_k \sim N(\mu, \gamma^2)$; the parameter $\gamma^2 \geq 0$ is called the interstudy variance. Then the estimated effects $\hat{\mu}_k$ for the respective studies are independent and approximately normal: $\hat{\mu}_k \sim N(\mu, w_k^{-1} + \gamma^2)$.

- Alternatively, let the transformed effects $\kappa_k \sim N(\kappa, \gamma^2)$. Then the estimated transformed effects $\hat{\kappa}_k$ for the respective studies are independent and approximately normal: $\hat{\kappa}_k \sim N(\kappa, 1/n_k + \gamma^2)$.

Questions

- What is the evidence against the null hypothesis of zero variance component $H_0 : \gamma^2 = 0$ (equivalent to the null hypothesis of homogeneity of study effects $\mu_k = \mu$ for all k, or homogeneity of transformed effects $\kappa_j = \kappa$ for all k) and for the alternative of a positive variance component $H_1 : \gamma^2 > 0$?

Test statistic and distribution

- If the hypotheses about the raw effects μ_k are of interest, the test statistic is the Cochran's Q defined by Equation (11.1).

- If the hypotheses about the transformed effects κ_k are of interest, the statistic Q^* is appropriate.

- Larger values of the test statistics Q or Q^* favor the alternative H_1 of the nonzero variance component over the null H_0 of homogeneity.

- For a fixed number of studies K and simultaneously growing sample sizes $n_k \to \infty$ (see Section 24.1.1 for details), the null distribution of Q or Q^* is approximately central χ^2_{K-1}.

- The distribution of Q or Q^* under alternatives $\gamma^2 > 0$ differs from the distribution under alternatives of heterogeneity of fixed effects; for equal sample sizes $n_k = N/K \to \infty$ and fixed K it is vescaled central chi-square distribution $(1 + \gamma^2 N/K)\chi^2_{K-1}$. Otherwise it is a quadratic form in normal random variables.

- For fixed sample sizes n_k and $K \to \infty$ the distribution of both Q and Q^* is approximately normal with moments given in Equation (24.11).

Transformation to evidence

- Let $M_r = \sum_k w_k^r$ be the sum of rth powers of the weights, and define $a = M_1 - M_2/M_1$ and $b = M_2 - 2M_3/M_1 + (M_2/M_1)^2$, $c = b/a^2$ and $d = c(K-1) - 1$.

- The evidence T'_Q in Cochran's Q (known weights) for the alternative $\gamma^2 > 0$ is defined by (24.12) as

$$T'_Q = \frac{1}{\sqrt{2c}} \left\{ \ln \left(\frac{c\,Q - d + \sqrt{d + (cQ - d)^2}}{1 + \sqrt{d+1}} \right) \right\}.$$

- The evidence T'_{Q^*} in Q^* for $\gamma^2 > 0$ is defined by (24.12) using the known weights $w_k = n_k$.

Interpretation

- The statistics Q, Q^* have the same null distributions under both fixed and random effects models. However, the alternatives themselves and the distributions of these statistics under alternatives differ under the fixed and random effects models. This results in differently defined evidence.

- In general, the evidences for heterogeneity in T'_Q and T'_{Q^*} for the random effects model are smaller than their fixed effects model counterparts T_Q and T_{Q^*}. Both increase with the number of studies K as the $\sqrt{K}$; but the evidence for fixed effects increases at the rate $\sqrt{n}$ for an average study size $n = N/K$ compared to the rate $\ln(n)$ for random effects. Therefore the evidence for random effects is unlikely to be large for small K regardless of study sample sizes.

- If the weights need to be estimated, then the moments (24.11) are only estimated, and therefore *vst* (24.12) may not be reliable. This extra source of variability caused by unknown weights undermines T_Q and T'_Q. Therefore the inference using T_{Q^*} and T'_{Q^*} which is based on transformed effects κ_k and uses known weights n_k is recommended.

11.2 Examples

11.2.1 Deaths by horse-kicks

We build on the analysis in Section 8.2.1 of the Bortkiewicz data on the numbers of soldiers killed per year by horse-kicks in each corps in the Prussian cavalry. The death rates over 20 years are denoted by μ_k for each of $K = 10$ corps. The observed death rates $\hat{\mu}_k$ and evidence T_k for the alternative $\mu > 0.5$ to the null $\mu \leq 0.5$ are listed in columns 3 and 6 of Table 8.1. The 'sample sizes' are the numbers of years of observation, $n_k = 20$ for all 10 corps.

Here the objective is to ascertain whether the death rates are homogeneous across the corps. Assume the estimated death rates satisfy $\hat{\mu}_k \sim N(\mu_k, \mu_k/n_k)$, so the estimated weights for Cochran's Q are $\hat{w}_k = n_k/\hat{\mu}_k$, and the estimated weighted mean is $\hat{\hat{\mu}}_w = 0.544$. Cochran's statistic for heterogeneity is $Q = 13.19$ with $p = 0.154$ found from the χ^2_9 distribution. The evidence for heterogeneity in Q is $T_Q = 0.960$, which is negligible.

The choice of $\mu_0 = 0.5$ in this example does not affect the homogeneity (or lack thereof) of the raw effects; the Q-test is location invariant. For the transformed effects $\kappa_k = 2(\sqrt{\mu_k} - \sqrt{\mu_0})$ of the Poisson model given by Equation (8.3) heterogeneity (or lack thereof) measured by Q^* is also free of the choice of μ_0, for the same reason.

The estimated transformed effects are $\hat{\kappa}_k = T_k/\sqrt{n_k}$ and their weighted mean is $\hat{\kappa} = \sum n_k\hat{\kappa}_k/\sum n_k = 0.122$, leading to $Q^* = 14.54$ with $p = 0.105$. The evidence in Q^* for heterogeneity of transformed effects $T_{Q^*} = 1.177$, slightly more than that in T_Q. One might expect them to be equal, given that the κ_k's are a monotonic function of the μ_k's, but Q and Q^* are different statistics. Generally speaking, Q approaches its limiting noncentral chi-squared distribution slower than does Q^* (see Chapter 24).

Looking at the contribution of each corps to the value of $Q_k^* = n_k(\hat{\kappa}_k - \hat{\kappa})^2$ it can be seen that corps XIV with the highest death rate contributes the largest term, 8.32, followed by corps VIII with the lowest death rate, contributing 2.43. Still, these values are not large enough to make the overall evidence for heterogeneity worth further consideration.

11.2.2 Drop in systolic blood pressure

We continue the analysis of the Mulrow *et al.* (2004) data given in Section 4.3.1. The objective is to assess the effectiveness of a weight-reducing diet for reducing systolic blood pressure in seven different studies. The average drop in systolic blood pressure for n_{2k} patients ($\bar{y}_k$) and n_{1k} controls ($\bar{x}_k$) and their pooled standard deviations $s_{\text{pool},k}$ were used to calculate the two-sample pooled t-statistics t_k with $\nu_k = n_{1k} + n_{2k} - 2$ degrees of freedom and the evidence for a positive effect T_k. All these values are shown in Table 11.1. In this example the raw effects of interest are differences in means $\mu_k = \mu_y - \mu_x$ estimated by $\hat{\mu}_k = \bar{y}_k - \bar{x}_k$. The standardized effects of

Table 11.1 Seven studies comparing drop in systolic blood pressure for treated patients undergoing a weight-loss regime from Section 4.3.1. Sample sizes n_{1k} and n_{2k}, the estimated effect $\hat{\mu}_k$, pooled sample standard deviation $s_{\text{pool},\,k}$, two-sample t-statistic $t_{\text{pool},\,k}$ and evidence for a positive effect T_k for each k are tabled along with contributions of each study to homogeneity statistics Q and Q^* denoted by Q_k and Q_k^*, respectively.

k	n_{1k}	n_{2k}	$\hat{\mu}_k$	$s_{\text{pool, k}}$	$t_{\text{pool},\,k}$	T_k	$\hat{w}_k$	Q_k	Q_k^*
1	24	27	−5.0	13.80	−1.29	−1.24	0.067	4.77	4.20
2	18	20	5.9	8.10	2.24	2.11	0.144	0.86	1.99
3	64	66	7.0	16.43	2.43	2.39	0.120	1.51	1.21
4	9	10	7.0	14.48	1.05	0.94	0.023	0.28	0.20
5	25	24	−7.0	18.51	−1.32	−1.27	0.036	3.91	4.25
6	5	5	7.3	6.18	1.87	1.42	0.065	0.97	1.13
7	14	19	2.6	6.34	1.16	1.09	0.201	0.15	0.19

interest, known as Cohen's d (Cohen 1988), are $\delta_k = \mu_k/\sigma_k$ for a common, to treatment and control groups, standard deviation σ_k estimated by the pooled standard deviation $s_{\text{pool},k}$.

To ascertain whether the differences in means are homogeneous across the studies, the estimated weights are $\hat{w}_k = n_{1k}n_{2k}/\{(n_{1k} + n_{2k})s^2_{\text{pooled},k}\}$, and the weighted mean is $\hat{\hat{\mu}}_w = 3.46$. Cochran's statistic for heterogeneity is $Q = 12.45$ with the p-value $p = 0.053$ found from the χ^2_6 distribution. The evidence in Q is $T_Q = 1.491$ which is weak.

The estimated transformed effects are $\hat{\kappa}_k = T_k/\sqrt{n_{1k} + n_{2k}}$ and their weighted mean is $\hat{\kappa} = \sum(n_{1k} + n_{2k})\hat{\kappa}_k/\sum(n_{1k} + n_{2k}) = 0.113$. This leads to $Q^* = 13.17$ with $p = 0.041$. The evidence in Q^* is $T_{Q^*} = 1.605$, which is weak evidence for heterogeneity of transformed effects. Looking at the contribution terms of each study to Q and Q^*, it is seen that studies 1 and 5 (the only two studies with negative results) make the main contributions to both statistics. In addition, study 2 makes a considerable contribution to Q^* but not to Q. This study has the second largest effect, and also a comparatively small standard deviation, resulting in a large standardized effect.

11.2.3 Effect of psychotherapy on hospital length of stay

We continue the analysis of Mumford *et al.* (1984) data introduced in Section 4.3.2. The objective is to compare the effectiveness of treatment 'psychotherapy' with control 'no therapy' for reducing length of stay (LOS) in hospital in days for eight different studies. The data are given in Table 4.2. The sample variances suggest that heteroscedasticity is present within most studies, so the Welch two-sample t-statistic is employed.

Entries from Tables 4.2 and 4.3 needed here are collected in Table 11.2. The row effects are the differences in mean LOS under two treatments, estimated by

Table 11.2 Statistical summaries of eight studies from Mumford *et al.* (1984) are listed in columns 2–8. The results compare mean difference in length of stay μ_k in hospital for patients receiving psychotherapy and no therapy. For each study k are given sample sizes n_{1k} and n_{2k}, the estimated effect $\hat{\mu}_k$, scale parameter $\hat{\sigma}_k$, standardized effect δ_k, Welch statistic $t_{\text{Welch},\,k}$ and evidence T_k for a positive δ_k.

k	n_{1k}	n_{2k}	$\hat{\mu}_k$	$\hat{\sigma}_k$	$\hat{\delta}_k$	$t_{\text{Welch,k}}$	T_k	$\hat{\kappa}_k$
1	13	13	−1.50	8.55	−0.175	−0.895	−0.86	−0.168
2	50	30	−1.20	4.03	−0.298	−2.662	−2.61	−0.129
3	35	35	−2.40	15.83	−0.152	−1.269	−1.24	−0.148
4	20	20	0.20	3.14	0.064	0.403	0.39	0.062
5	10	10	0.18	1.72	0.105	0.469	0.44	0.099
6	14	13	−0.60	2.02	−0.297	−1.544	−1.47	−0.283
7	9	9	−2.22	3.31	−0.671	−2.845	−2.53	−0.598
8	8	8	−0.88	2.26	−0.388	−1.554	−1.41	−0.352

$\hat{\mu}_k = \bar{x}_{2k} - \bar{x}_{1k}$. The standard errors of these estimates are $\hat{\sigma}_k/\sqrt{N_k}$, where $N_k = n_{1k} + n_{2k}$.

The estimated standardized effects are $\hat{\delta}_k = \hat{\mu}_k/\hat{\sigma}_k$. The estimated transformed effects $\hat{\kappa}_k$ are $\hat{\kappa}_k = N_k^{-1/2}\,T_k$, and are listed in the last column of Table 11.2. Six of the eight studies yield negative effects which suggest that psychotherapy does make a difference. In this discussion we wish to test whether the effects (raw and/or standardized effects) are homogeneous.

Cochran's statistic for heterogeneity is $Q = 14.204$ with p-value 0.048 found from the χ^2_7 distribution. The evidence for heterogeneity of fixed effects in Q is $T_Q = 1.54$, which is very weak.

The statistic for heterogeneity of the transformed effects is $Q^* = 8.814$ with p-value 0.266. The evidence in Q^* is 0.594. There is negligible evidence for heterogeneity of transformed effects. This happens because comparatively large standardized effects in the last three studies correspond to small sample sizes, and therefore do not contribute much to Q^*.

11.2.4 Diuretics in pregnancy and risk of pre-eclamsia

We continue the analysis of Collins *et al.* (1985) data from Section 7.2.2. The objective is to investigate the possible benefit of taking diuretics during pregnancy to prevent pre-eclamsia. The raw effects are the differences in absolute risk of pre-eclamsia in nine clinical trials of n_{2k} patients and n_{1k} controls, $k = 1, \cdots, 9$. The total sample sizes are $N_k = n_{1k} + n_{2k}$. The standardized effects of interest are correlation effect sizes ρ_k. The evidence for a positive effect T_k was calculated from the *vst* (19.1). The data, correlation effect sizes $\hat{\rho}_k$ and estimated transformed effects $\hat{\kappa}_k$ are given in Table 7.1. Here we calculate evidence for heterogeneity of fixed transformed effects, and also the evidence for random transformed effects.

The statistic for heterogeneity of the transformed effects is $Q^* = 22.4$, and the evidence for heterogeneity in Q^* is almost moderate at $T_{Q^*} = 2.41$. The constants required for calculation of the evidence for random transformed effects given by (24.12) are $c = 0.268$ and $d = 1.145$. The value of $T'_{Q^*} = 1.900$. Thus there is a weak evidence for random transformed effects. As expected, the evidence for heterogeneity of random effects is weaker than that for fixed effects, but in this example the random transformed effects model is a reasonable way forward, whereas combining very heterogeneous effects through the fixed equal effects model is rather foolhardy.

By contrast, there is no need to even calculate the evidence for random effects in any of the previous examples in this section; it would be even weaker than the weak to negligible evidence we found for the heterogeneity of fixed effects which was calculated. Another consideration in adopting a random transformed effects model is whether the K studies can reasonably be viewed as a random sample of studies from a larger population of studies.

12

Combining evidence: fixed standardized effects model

In the previous chapter quantitative methods for helping to decide whether to choose a fixed or random standardized effects model are provided. Therefore it is assumed here that the researcher has already decided to adopt a fixed standardized effects model.

The choice between 'fixed and equal' and 'fixed but unequal' standardized effects models is aided by the evidence T_{Q^*} in Q^* for heterogeneity of the transformed standardized effects $\hat{\kappa}_k$, which indirectly measures the heterogeneity of the standardized effects $\hat{\delta}_k$, because all $\kappa_k = \mathcal{K}(\delta_k)$, and $\mathcal{K}$ is a monotonically increasing function assumed common to all models.

The distinction between the fixed and equal and fixed but unequal standardized effects models is conceptually important but the methodology is exactly the same. In the first model all $\delta_k = \delta$ are assumed equal and of course δ is the parameter of interest; in the second model, the δ_k are combined by $\delta = \mathcal{K}^{-1}(\kappa)$, where κ is a weighted mean (weights equal to the sample sizes) of the κ_k's. The theory is given in Chapter 25. Other options for the heterogeneous case are presented in Chapters 13 and 14.

The standard meta-analytic approach is to first carry out a Cochran Q-test for homogeneity of raw effects and, if it is not significant, combine the effects from the respective studies using a weighted mean, with estimated inverse variance weights. The same approach can be adapted to standardized effects. These methods are also illustrated here for the sake of comparison and completeness. Their theory is well established and available in many books on meta analysis from Hedges and Olkin (1985) to Sutton *et al.* (2000).

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

12.1 Methodology

Data and model

- Given K studies of sizes n_k measuring potentially different effects μ_k, for $k = 1, \dots, K$.
- The estimated effects $\hat{\mu}_k$, $k = 1, \dots, K$, for the respective studies are mutually independent and approximately normal with means $\hat{\mu}_k$ and variances w_k^{-1} estimated by $\hat{w}_k^{-1}$. A representative μ for the K studies could be $\bar{\mu}_w = \sum_k w_k \mu_k / \sum_j w_j$.
- Standardized effects are denoted by δ_k, and their transformed versions by $\kappa_k = \mathcal{K}(\delta_k)$. Their weighted mean is $\kappa = \sum_k n_k \kappa_k / N$, where $N = \sum n_k$ is the total sample size.
- Evidence in the kth study for $\delta_k > 0$ is T_k which is approximately distributed $N(\sqrt{n_k}\,\kappa_k, 1)$. The transformed effects are estimated by $\hat{\kappa}_k = T_k/\sqrt{n_k}$; the estimate $\hat{\kappa}_k$ is approximately normal with mean κ_k and variance n_k^{-1}.

Questions

- What is an estimate of a representative effect μ and a confidence interval for μ?
- How does one define a representative standardized effect δ for the K studies, without assuming that the δ_k's are equal?
- What is the evidence for $\delta > 0$?
- What is a confidence interval for such a δ?

Transformation to evidence

- A representative κ for the K studies is the weighted mean $\kappa = \sum_k n_k \kappa_k / N$; and the representative $\delta = \mathcal{K}^{-1}(\kappa)$. If it turns out that all δ_k are equal, this δ equals the common value, because $\kappa_k = \mathcal{K}(\delta_k)$.
- Given independent $(T_1, \dots, T_K)$, the *combined evidence* for $\delta > 0$ in the K studies is

$$T_{1:K} = \frac{\sqrt{n_1}\,T_1 + \cdots + \sqrt{n_K}\,T_K}{\sqrt{n_1 + \cdots + n_K}} \,. \tag{12.1}$$

As usual, when $T_{1:K}$ is negative, its magnitude $|T_{1:K}|$ is interpreted as evidence for $\delta < 0$. Because of the properties of the individual T_k's, the combined evidence is approximately normal with mean $\mathrm{E}[T_{1:K}] \doteq \sqrt{N}\,\kappa$ and variance 1. For further discussion of this definition, see Section 25.2.2.

Interpretation

- The combined evidence $T_{1:K}$ lies on the calibration scale, so can be interpreted as an estimator of its expectation with a standard normal error. It allows for cancellation of positive and negative evidence from conflicting studies.

Confidence intervals

- The weighted mean of effects $\bar{\mu}_w = \sum_k w_k \mu_k / \sum_k w_k$ is traditionally estimated by $\hat{\bar{\mu}}_w = \sum_k \hat{w}_k \hat{\mu}_k / \sum_k \hat{w}_k$; and the $100(1-\alpha)$ % confidence interval for $\bar{\mu}_w$ has endpoints $\hat{\bar{\mu}}_w \pm z_{1-\alpha/2} / \sqrt{\sum_k \hat{w}_k}$. If the effects are equal to μ, say, then of course this is a confidence interval for $\mu = \bar{\mu}_w$.
- A $100(1-\alpha)$ % confidence interval for κ has endpoints defined by $(T_{1:K} \pm z_{1-\alpha/2})/\sqrt{N}$. An interval of the same confidence for $\delta = \mathcal{K}^{-1}(\kappa)$ is obtained by applying $\mathcal{K}^{-1}$ to these endpoints. If the standardized effects are equal to δ, say, then it is a confidence interval for the common δ, rather than the representative δ that transforms into the weighted average of the κ_k's, defined earlier.
- When the effects μ_k are one-to-one functions of the standardized effects δ_k, two different interval estimates for the representative effect can be obtained from the above two approaches. The interval based on the transformed standardized effects technique has, generally speaking, better coverage properties due to the variance stabilization process, which also improves the normal approximation.

Extension required for nuisance parameters

- When the Key $\mathcal{K}$ depends not only on a standardized effect δ but also on a nuisance parameter ξ, the standardized effects are $\kappa_k = \mathcal{K}(\delta_k, \xi_k)$.
- If the Key is strictly monotonic in both arguments, a representative ξ can be defined. The choice $\xi = \sum n_k \xi_k / N$ seems reasonable; for more discussion, see Section 25.2.3.
- Once a representative ξ is defined, its estimate $\hat{\xi}$ can be used to solve the equation $\mathcal{K}(\hat{\delta}, \hat{\xi}) = \hat{\kappa}$ for $\hat{\delta}$.
- Endpoints of the level $1-\alpha$ confidence interval for $\delta = \mathcal{K}^{-1}(\kappa, \xi)$ are solutions for $\hat{\delta}$ to the equation $\mathcal{K}(\hat{\delta}, \hat{\xi}) = \hat{\kappa} \pm z_{1-\alpha/2}/\sqrt{N}$.

12.2 Examples

12.2.1 Deaths by horse-kicks

We continue with the analysis of these data and model from Section 8.2.1 and the test for heterogeneity in Section 11.2.1. Recall that for each of the 10 cavalry corps in the

Prussian army, the numbers of deaths by horse-kicks were modeled by the Poisson distribution with respective rates μ_k, and the evidences T_k for $\mu_k > 0.5$ found for each corps are listed in Table 8.1. Further, we found negligible evidence for heterogeneity of effects and similarly negligible evidence for heterogeneity of transformed effects, so we adopt the fixed and equal standardized effects model. The representative κ is related to the common mean μ by $\kappa = \mathcal{K}(\mu|\mu_0) = 2(\sqrt{\mu} - \sqrt{\mu_0})$, where $\mu_0 = 0.5$.

The evidence for $\kappa > 0$ is given by Equation (12.1), and for equal sample sizes of $n_k = 20$ reduces to $T_{1:10} = \sum_{k=1}^{10} T_k/\sqrt{10} = 1.73$, which is weak evidence for noncompliance of regulations $\mu > 0.5$. Note that this is slightly less than the weak evidence $T_{200} = 2.09$ for $\mu > 0.5$ obtained in Section 11.2.1. This latter measure of evidence utilized the fact that all 10 cavalry corps could be considered as a whole, with the same Poisson model. This example suggests that it is better to combine all the K study test statistics before carrying out a single variance stabilization to obtain evidence, rather than use a *vst* for each and then combine the evidence. However, the former option will not be available for most models.

Another objective is to find interval estimates of the representative standardized effect δ. But for the one-sample Poisson model $\delta = (\mu - \mu_0)/\sqrt{\mu_0}$ and we are more interested in an interval for μ. An inverse transformation to find μ is given by Equation (8.4) as $\mathcal{K}^{-1}(y|\mu_0) = (\max((y/2 + \sqrt{\mu_0}), 0))^2$. The estimated transformed effects $\hat{\kappa}_k = T_k/\sqrt{n_k}$ and their weighted mean $\hat{\kappa} = \sum n_k\hat{\kappa}_k / \sum N_k = 0.122$ were calculated in Section 11.2.1. The estimated death rate $\hat{\mu} = \mathcal{K}^{-1}(0.122) = 0.506$, and the 95 % confidence interval for the death rate is [0.489, 0.701]; it is found by applying $\mathcal{K}^{-1}(y|\mu_0)$ to the endpoints of the interval for κ which are $(T_{1:10} \pm 1.96)/\sqrt{200}$.

The standard meta-analytic estimates are $\hat{\hat{\mu}}_w = 0.544$ and the confidence interval $\hat{\hat{\mu}}_w \pm z_{1-\alpha/2}/\sqrt{\sum w_k} = [0.442, 0.646]$. Interestingly, $\hat{\mu} < \hat{\hat{\mu}}_w$, but the corresponding confidence interval is more to the right. The first interval should have better coverage, as follows from discussions in Sections 17.3.5 and 17.3.6.

12.2.2 Drop in systolic blood pressure

We build on the analysis of the Mulrow *et al.* (2004) data in Section 11.2.2. The objective is to assess the effectiveness of a weight-reducing diet for lowering systolic blood pressure in seven different studies. The estimated raw effects $\hat{\mu}_k$ are the differences in average drop in systolic blood pressure for n_{2k} patients $(\bar{y}_k)$ and n_{1k} controls $(\bar{x}_k)$. The sample sizes are $N_k = n_{1k} + n_{2k}$, and the standardized effects of interest are Cohen's $d_k = \mu_k/\sigma_k$, where a common unknown standard deviation σ_k is estimated by the pooled standard deviations $s_{\text{pool},k}$ (see Section 4.1). The evidence T_k for a positive standardized effect $d_k > 0$ is based on the two-sample pooled t-statistics t_k, with all results given in Table 11.1.

The estimated transformed effects $\hat{\kappa}_k = T_k/\sqrt{N_k}$ and their weighted mean $\hat{\kappa} = \sum N_k\hat{\kappa}_k / \sum N_k = 0.113$ were calculated in Section 11.2.2, along with the evidence for heterogeneity of $T_{Q^*} = 1.605$. So there is only marginal to weak evidence for heterogeneity of transformed effects. We adopt the fixed but unequal standardized effects model. This means that we need to define a representative standardized effect δ.

For each study k the transformed effect is $\kappa_k = \mathcal{K}(\sqrt{q_k(1-q_k)}\, d_k)$, where $\mathcal{K}(x) = \sqrt{2}\,\sinh^{-1}(x/\sqrt{2})$ and $q_k = n_k/N_k$. Here the Key is a function of both d_k and the study-specific constant q_k (see Section 4.1). Therefore a representative value of q should be chosen prior to solving the equation $\kappa = \mathcal{K}(\sqrt{q(1-q)}\,\delta)$ for a representative standardized effect δ. Let $q = \sum N_k q_k/N = \sum n_k/N = n/N$, so that q is the overall proportion of patients undergoing the weight-reducing diet. Then

$$\delta = \delta_q(\kappa) = \frac{\sqrt{2}\,\sinh(\kappa/\sqrt{2})}{\sqrt{q(1-q)}} \,.$$

An estimate for δ is $\hat{\delta} = \delta_q(\hat{\kappa})$. A confidence interval for κ has endpoints $(\hat{\kappa} \pm z_{1-\alpha/2})/\sqrt{N}$. Applying the function δ_q to this interval yields an interval for δ with the same confidence coefficient.

The evidence for $\delta > 0$ is defined in (12.1) and equal to $T_{1:7} = 2.06$.

A point estimate of δ is $\hat{\delta} = 0.227$ with the 95 % confidence interval $[0.068, 0.386]$. Note that the confidence interval is not quite symmetric around $\hat{\delta}$: the lower limit is $\hat{\delta} - 0.1588$, and the upper limit is $\hat{\delta} + 0.1595$. This reflects the skewness of the noncentral t-distribution.

13

Combining evidence: random standardized effects model

In Chapter 11 quantitative methods for deciding whether to choose a fixed or random standardized effects model are discussed, so it is assumed here that the researcher has already decided to adopt a random standardized effects model. Thus there is good reason to suppose an interstudy variance component $\gamma^2 > 0$ exists and must be accounted for. While we discuss two estimates for γ, both are biased upwards for small γ, and small γ seems to be the rule, not the exception, in applications. Fortunately, it is not necessary to estimate γ to find evidence for a positive standardized effect δ, or to find interval estimates for δ in the presence of γ. The theory is given in Section 25.3. For other options to proceed in the case of heterogeneous effects, see the fixed effects model in Chapter 12 and the meta-regression in Chapter 14.

13.1 Methodology

Data and model

- For each of K studies adopting the same model, evidence T_k is available in the kth study for a positive standardized effect $\delta_k > 0$. Further, for the same Key $\mathcal{K}$ common to all the studies, the transformed standardized effects are defined by $\kappa_k = \mathcal{K}(\delta_k)$.

- For the *fixed* standardized effects model, $T_k \sim N(\sqrt{n_k}\,\kappa_k, 1)$, and the transformed effects are estimated by $\hat{\kappa}_k = T_k/\sqrt{n_k}$, which are approximately distributed $N(\kappa_k, n_k^{-1})$. Here the κ_k's are constants.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

- For the *random* standardized effects model, the κ_k's are assumed to be a random sample from the $N(\kappa, \gamma^2)$ model, where κ, γ^2 are unknown; and the just described distributions for the fixed model are now considered to be conditional distributions, with the kth conditional on the value of κ_k. It follows that the unconditional distribution of $\hat{\kappa}_k$ is approximately $N(\kappa, n_k^{-1} + \gamma^2)$. The parameters of interest are κ and the representative standardized effect $\delta = \mathcal{K}^{-1}(\kappa)$. For more discussion of this model, see Section 25.3.

Questions

- What are point estimates of the mean transformed effect κ and of the interstudy variance component γ^2?
- What is the evidence for $\kappa > 0$?
- What is a confidence interval for κ and for $\delta = \mathcal{K}^{-1}(\kappa)$?

Point estimates of the mean transformed effect κ and $\delta = \mathcal{K}^{-1}(\kappa)$

- Let $\bar{\kappa} = (\sum_k \hat{\kappa}_k)/K$ and $s_\kappa^2 = \sum_k (\hat{\kappa}_k - \bar{\kappa})^2/(K-1)$ denote the sample mean and variance of the $\hat{\kappa}_k$'s.
- Clearly $\bar{\kappa}$ is an unbiased estimator of κ, with variance $\text{Var}[\bar{\kappa}] = \sigma^2/K$, where

$$\sigma^2 = \gamma^2 + \frac{1}{K}\sum_k \frac{1}{n_k}.$$

  It is left to the reader to show that $\text{E}[s_\kappa^2] = \sigma^2$, so s_κ^2 is an unbiased estimator of σ^2, and one can estimate σ^2 without estimating γ^2. Thus approximately $\bar{\kappa} \sim N(\kappa, \text{E}[s_\kappa^2]/K)$, and the standard error of estimation is $\text{SE}[\bar{\kappa}] = s_\kappa/\sqrt{K}$.

- The standardized effect $\delta = \mathcal{K}^{-1}(\kappa)$ is estimated by $\bar{\delta} = \mathcal{K}^{-1}(\bar{\kappa})$.
- An alternative unbiased estimator of κ employed for the case of fixed effects in Chapter 12 is $\hat{\kappa} = \sum n_k \hat{\kappa}_k/N$. For this model its distribution is $N(\kappa, N^{-1}(1 + \gamma^2 \sum n_k^2/N))$. The two estimators $\bar{\kappa}$ and $\hat{\kappa}$ coincide when the sample sizes are equal $n_k = n = N/K$.

Test statistic and transformation to evidence

- First note that when the sample sizes are all equal to n, the κ_k's are just a random sample from $N(\kappa, \sigma^2)$ with $\sigma^2 = \gamma^2 + 1/n$. It follows that $\sqrt{K}\,\bar{\kappa}/s_\kappa \sim t_{K-1}(\lambda)$, with noncentrality parameter $\lambda = \sqrt{K}\,\kappa/\sigma$. This noncentral Student t-distribution can also be useful when the sample sizes are large enough so that their reciprocals are small compared to γ^2.
- The test statistic for an alternative $\kappa > 0$ is $S_{K-1} = \sqrt{K}\,\bar{\kappa}/s_\kappa$.

- The evidence for $\kappa > 0$ and hence $\delta > 0$ is given by

$$T^*_{1:K} = \sqrt{2K}\ \sinh^{-1}\left(\frac{\bar{\kappa}}{\sqrt{2}\, s_\kappa}\right).$$

Interpretation

- When all sample sizes are equal, this $T^*_{1:K}$ is approximately normal with variance 1 and mean $\mathrm{E}[T^*_{1:K}] \doteq \sqrt{2K}\ \sinh^{-1}(\kappa/\sqrt{2}\,\sigma)$. This mean evidence is monotonically increasing in κ and monotonically decreasing in γ, because $\sigma^2 = 1/n + \gamma^2$. Thus a large number of studies K will be necessary to find even weak evidence for $\kappa > 0$ when γ is large.
- When all sample sizes are approximately equal, or all their reciprocals small compared to γ^2, the above results are expected to still be applicable, because simulations of t-intervals for κ found under these conditions demonstrate good coverage probabilities.

Confidence intervals for mean transformed effect κ and $\delta = \mathcal{K}^{-1}(\kappa)$

- For equal sample sizes $n_k = N/K$ a nominal $100(1-\alpha)$ % confidence interval for κ has endpoints $[L, U]$ defined by $\bar{\kappa} \pm t_{K-1,1-\alpha/2}\sqrt{s^2_\kappa/K}$.
- $[\mathcal{K}^{-1}(L), \mathcal{K}^{-1}(U)]$ covers δ with the same confidence coefficient.
- The above intervals are approximately valid for any general sample sizes n_k when they are large enough so that the values $1/n_k$ are small relative to γ^2 (see Section 25.3).
- In Section 25.3 it is shown that confidence intervals based on $\hat{\kappa}$ are not reliable under the random transformed effects model, so are not recommended.

Estimates of the parameter γ^2

- Recall the Cochran statistic $Q^* = \sum_k n_k(\hat{\kappa}_k - \hat{\kappa})^2$ of Equation (11.2) for assessing heterogeneity. Its expectation can be used to derive the method of moments estimator $\hat{\gamma}^2_M = (Q^* - (k-1))/(N - \sum n^2_k/N)$, but this can take on negative values. DerSimonian and Laird (1986) proposed the modified estimator $\hat{\gamma}^2_{DL} = \max\{0, \hat{\gamma}^2_M\}$ in order to correct this problem.
- Variance $\mathrm{Var}(\hat{\gamma}^2_M) = \mathrm{Var}(Q^*)/(N - \sum n^2_k/N)^2$, where $\mathrm{Var}(Q^*)$ is given in Equation (24.11). The variance is small only when the number of studies K is large.
- Another estimator γ^2 is $\hat{\gamma}^2_S = \max\{0,\ s^2_\kappa - \frac{1}{K}\sum_k \frac{1}{n_k}\}$. Once more, the variance of $\hat{\gamma}^2_S$ is small only for large K. Simulations show that both estimators are biased upwards for small γ^2.

13.2 Example

13.2.1 Diuretics in pregnancy and risk of pre-eclamsia

We continue the analysis of the Collins *et al.* (1985) data studied in Sections 7.2.2 and 11.2.4. The objective is to investigate the benefits of taking diuretics during pregnancy on the risk of pre-eclamsia by combining the evidence from nine clinical trials. The raw effects are the differences in absolute risk $\Delta_k = p_{1k} - p_{2k}$ for n_{2k} patients and n_{1k} controls, $k = 1, \ldots, 9$. The sample sizes are $N_k = n_{1k} + n_{2k}$. The standardized effects of interest are correlation effect sizes $\rho_k = \Delta_k/\sqrt{\zeta_k}$, where ζ_k are study-specific parameters defined by $\zeta_k = \{p_k(1 - p_k)\}/\{q_k(1 - q_k)\}$ for $q_k = n_{2k}/(n_{1k} + n_{2k})$, and $p_k = q_k p_{1k} + (1 - q_k)p_{2k}$. The evidence for a positive effect T_k was calculated from the *vst* (19.1). The data, correlation effect sizes $\hat{\rho}$ and estimated transformed effects $\hat{\kappa}$ are given in Table 7.1. There is a weak evidence $T'_{Q^*} = 1.900$ for random transformed effects calculated in Section 11.2.4. The random effects model is used here to combine the evidence for a positive representative correlation effect ρ.

The statistic $S_{K-1} = 1.796$ and the p-value of 0.055 when performing a conventional t-test for $\kappa > 0$ may be found from central t_8-distribution. The evidence $T^*_{[1:K]} = 1.749$ seems larger than it should be when compared to the p-value until we recall that the evidence in the t-test (or any other evidence) is not routinely calibrated to provide a value of 1.65 when $p = 0.05$ (see Section 20.4.1 for discussion). If such a calibration were desired, a corrected evidence $T_{\text{corrected}} = 1.596$ is calculated as $(1 - 0.7/(K - 1))\sqrt{2K}\ \sinh^{-1}(S_{K-1}/\sqrt{2K})$, as suggested in Section 20.4.1. In any case there is a weak evidence of a positive correlation effect, so the risk of pre-eclamsia may be reduced by diuretics. The point estimate of transformed effect is $\bar{\kappa} = 0.079$ and the 95 % confidence interval for κ is $(-0.022, 0.181)$. Finally, point and interval estimates of the correlation effect ρ are $\bar{\rho} = 0.079$ with the 95 % confidence interval $(-0.022, 0.180)$.

It would be very easy to calculate point and interval estimates of the standardized effect δ, since $\delta^2 = \rho^2/(1 - \rho^2)$, but it is not straightforward to estimate representative absolute risk difference Δ. To do that a reasonable common value of ζ is needed, and there is no evident way to define such a value.

The estimates of interstudy variance are rather different: $\hat{\gamma}_{DL} = 0.003$ and $\hat{\gamma}_S = 0.013$. Since the number of studies $K = 9$ is not large, the variation of these estimates is rather high, and the confidence intervals for ρ and κ given above are also rather wide.

For comparative purposes, the same evidence was combined under fixed transformed effects model, even though this model may be wrong to use due to high heterogeneity ($Q^* = 22.4$ and $T_{Q^*} = 2.41$). The values are $\hat{\rho} = 0.057$, with the 95 % confidence interval $(0.034, 0.081)$. This interval is considerably more narrow than the interval for ρ under random effects.

14

Meta-regression

In a meta analysis results from several studies are combined. When the studies are heterogeneous, straightforward combination of test results may be too simplistic and more sophisticasted techniques should be used. One such technique is meta-regression. In this model, the effect sizes estimated in the individual studies are modeled as functions of one or more characteristics of the studies (see Thompson and Higgins 2002). The meta-regression model (fixed effects regression) is an extension of the fixed effects model and is most appropriate when all variation above and beyond the sampling error between study outcomes can be accounted for by the covariates included. A mixed model is more suitable when the covariates explain only part of the variation, and a random effect term is used to account for a remainder (Sutton *et al.* 2000, Chapter 6). Only fixed effects regression is considered in this chapter.

Because the transformation of the study outcomes to evidences by applying an appropriate *vst* simplifies the distributional properties, it is easier to formulate an accurate model.

14.1 Methodology

Data and model

- The basic data consist of the observed evidence $T_1, \ldots, T_K$ from K studies of sample sizes n_j, $j = 1, \ldots, K$.

- Let $Y_j = \hat{\kappa}_j = n_j^{-1/2} T_j \sim N(\mathcal{K}(\delta_j), n_j^{-1})$ be the estimated transformed effects, which are related to the (standardized or raw) effects δ_j as indicated.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte
© 2008 John Wiley & Sons, Ltd

- In addition, we are given $u < K$ predictor variables $X = (X_1, \ldots, X_u)$, all of which are study characteristics.

Questions

- What is the evidence that the covariates, which describe the characteristics or circumstances of the studies, are related to the effect sizes δ?
- To make the relationship between the covariates and the raw effect linear, one has to apply a linearizing transformation $f(\delta)$, which is usually treated as known. The model then only depends on a u-vector of regression coefficients $(\beta_1, \ldots, \beta_u)$ and is of the form $f(\delta) = \beta_1 X_1 + \cdots + \beta_u X_u$. In this model, the regression coefficients relate directly to the raw effects δ.
- The transformed effects $\kappa_k = \mathcal{K}(\delta_k)$ have estimates with approximately normal distributions. No additional linearization is needed. This suggests the alternative model $\kappa = \beta_1 X_1 + \cdots + \beta_u X_u$, in which the regression coefficients are directly related to the transformed effects. This corresponds to a nonlinear model for the raw effects, $\delta = \mathcal{K}^{-1}(\beta_1 X_1 + \cdots + \beta_u X_u)$. The two models are equal when $f(\cdot)$ equals $\mathcal{K}(\cdot)$.
- What are the estimates and confidence intervals for the regression coefficients $\beta_1, \ldots, \beta_u$?

Theory

- In the first model, which involves a linearizing transformation, we propose the model
$$Y_k = \mathcal{K}(f^{-1}(\beta_1 X_{k1} + \cdots + \beta_u X_{ku})) + \epsilon_k, \tag{14.1}$$
where we assume that $\epsilon_1, \ldots, \epsilon_k$ are independent with $\epsilon_k \sim N(0, n_k^{-1})$. This choice is justified by the fact that the estimated evidences are roughly normally distributed.
- Relationship (14.1) is a nonlinear regression model with known variances. It is also part of the family of generalized linear models (GLMs) with the *link function* $g(y) = f(\mathcal{K}^{-1}(y))$ (see McCullagh and Nelder, 1999, for the general theory of GLMs).
- The simpler model is a weighted linear regression model with known weights n_k, because the response variable Y has an expectation that is linear in the covariates and a known variance of $1/n_k$
$$Y_k = \beta_1 X_{k1} + \cdots + \beta_u X_{ku} + \epsilon_k. \tag{14.2}$$
The two models (14.1) and (14.2) are equivalent when $\mathcal{K}(\cdot) = f(\cdot)$.

Using standard software

- The GLM analysis can be performed in a number of statistical software packages, including R, SPLUS, SAS among others. Usually the software has a list of ready-made link functions, corresponding to particular families of distributions, but the functions of interest to us are not included. It is, however, possible to pass the required information in the form of the link function $g(y)$, its inverse $g^{-1}(\cdot)$ and its derivative $g'(\cdot)$. Details differ for different software packages. One may also need to specify a variance function, the description of the variance as a function of mean. Since the errors in (14.1) are normally distributed, the variance function is constant. Most statistical packages also include routines for nonlinear regressions, which is an alternative way of fitting (14.1).
- The linear regression model relating Y linearly to the covariates (14.2) is easiest to fit. Virtually all statistical packages include a least squares regression solver.
- In all cases the sample sizes n_k must be used as case weights.
- The software for GLMs, nonlinear and linear regressions will compute a value for the squared global scale parameter σ^2 and make use of it in computing standard errors and confidence intervals. In our models, this global scale is known to be equal to one.

What to look for in the output

- The estimates $\hat{\beta}_1, \ldots, \hat{\beta}_u$ of coefficients are obtained directly from the output.
- The standard errors of the estimates of the regression coefficients from the output should be divided by the estimated global scale $\hat{\sigma}$ to account for the fact that the global scale is one. These corrected values of standard errors are denoted by s.e. $[\hat{\beta}_k]$ in what follows.

Tests and confidence intervals

- The t-tests for coefficients $\beta_k \neq 0$ given in the output should be changed to z-tests based on the values of $z = \hat{\sigma} t$. These can easily be transformed to two-sided evidence.
- A confidence interval for the coefficient β_k is given by $[\hat{\beta}_k \pm \text{s.e.}\ [\hat{\beta}_k] z_{1-\alpha/2}]$.
- The weighted residual sums of squares for regression or so called deviances for GLMs are χ^2_{K-u}-distributed. They can be used for lack-of-fit testing, and transformed to evidence for lack-of-fit via Equation (22.1).
- Suppose a model $H_v \subset H_u$ has only v parameters $\beta_1, \ldots, \beta_v$ whereas a model H_u includes $u > v$ parameters $\beta_1, \ldots, \beta_v, \beta_{v+1}, \ldots, \beta_u$. The difference of weighted residual sums of squares (or deviances in the case of GLMs) has the χ^2_{u-v} distribution, and can be used to test $\beta_{v+1} = \ldots = \beta_u = 0$. Large values indicate nonzero coefficients, i.e. the lack of fit of model H_v as compared to H_u.

Traditional meta-regression

- Model (14.1) is a counterpart of the weighted regression for effects traditionally used in meta analysis. This model assumes a normal distribution for the linearized effect sizes $f(\delta)$ and uses weights equal to the inverse estimated variances of $\widehat{f(\delta)}$. Sometimes logistic regression is used when the effects of interest are odds ratios. See Sutton *et al.* (2000, Chapter 6) for more details.
- In the model (14.1) the weights are known sample sizes, not estimated variances. The results are thus more stable. In addition, the assumption of normality is better justified due to the *vst* that was applied when computing evidence and prior to the meta modeling.

14.2 Commonly encountered situations

In this section we discuss some common situations in which meta-regression is used. Consider a meta analysis of K two-sample studies (treatment versus control) of sizes $N_k = n_{1k} + n_{2k}$, respectively. Throughout the remainder of this chapter, the subscript k which ranges over the studies, is often suppressed for simplicity of notation. The additional index is 1 for control and 2 for treatment. Denote by $q_k = n_{2k}/N_k$ the proportion of observations in the treatment arm of a study.

We will consider three common types of effects in this section. First, Cohen's standardized effect $d_{\text{Cohen}} = (\mu_T - \mu_C)/\sigma$, next the difference $\Delta = p_T - p_C$ of two binomial proportions and finally the relative risk $\rho = \mu_T/\mu_C$ for two Poisson rates.

14.2.1 Standardized difference of means

When the outcome of interest is a continuous variable and the variances σ^2 are assumed to be equal between the two arms of a study, the results are usually reported as standardized differences of the means for the two arms of the study $\hat{d}_{\text{Cohen}} = (\bar{x}_2 - \bar{x}_1)/s_{\text{pooled}}$. In this expression, $\bar{x}_2$ and $\bar{x}_1$ are the sample means, and s_{pooled} is pooled standard deviation. This statistic is an estimate of $d_{\text{Cohen}} = (\mu_2 - \mu_1)/\sigma$ (see Cohen, 1988).

In Section 4.1 the comparison of two group means is discussed. The standardized effect size is $\delta = (q(1-q))^{1/2}(\mu_2 - \mu_1)/\sigma$, where $q = n_2/(n_1 + n_2)$. The corresponding two-sample t-test statistic is $t_{\text{pooled}} = \sqrt{N}\hat{\delta}$. This has a t-distribution with $\nu = N - 2$ degrees of freedom and the corresponding *vst* is Azorin's (1953) transformation with the key

$$\mathcal{K}(\delta) = \sqrt{2}\sinh^{-1}(\delta/\sqrt{2}) = \sqrt{2}\ln(\delta/\sqrt{2} + \sqrt{1 + \delta^2/2}).$$

To perform a GLM or nonlinear least squares meta-regression the linear relationship between the covariates and the effect sizes is assumed to hold for $\hat{d} = (q(1-q))^{-1/2}\hat{\delta}$. The link function is thus $g(y) = (q(1-q))^{-1/2}\mathcal{K}^{-1}(y)$, where

$$\mathcal{K}^{-1}(y) = \sqrt{2}\sinh(y/\sqrt{2}) = \{\exp(y/\sqrt{2}) - \exp(-y/\sqrt{2})\}/\sqrt{2}.$$

The inverse link function is

$$g^{-1}(x) = \mathcal{K}((q(1-q))^{1/2}x) = \sqrt{2}\sinh^{-1}((q(1-q))^{1/2}x/\sqrt{2}),$$

and its derivative is

$$g'(y) = (q(1-q))^{-1/2}(\exp(y/\sqrt{2}) + \exp(-y/\sqrt{2}))/2.$$

When the variances in the two arms of a study are not assumed to be equal, the Welch t-test should be used instead of the t-test with pooled variances, and the appropriate *vst* may be found in Section 4.2.

14.2.2 Difference in risk (two binomial proportions)

Consider a linear model for a difference in risk in a meta analysis of K studies. The risk in question may be a risk of a disease and a treatment under consideration may be a medical intervention or a behavioral change, such as a smoking cessation program. The number of cases in the treatment and control arms (X_2 or X_1 respectively) can be modeled as binomial random variables, and the difference in risk is the difference of binomial proportions $\Delta = p_2 - p_1$ between the two arms of the study.

This basic model is considered in Chapter 19. As usual, $q = n_2/(n_1 + n_2)$. The parameter $p = qp_1 + (1-q)p_2$ and function $\zeta = \{p(1-p)\}/\{q(1-q)\}$ introduced in Section 19.1.1, are needed to specify the key function, which is given by (see Equation (19.1))

$$\mathcal{K}(\Delta) = \arcsin(\Delta/\sqrt{\zeta}). \tag{14.3}$$

Because the range of Δ is restricted, various choices for the linearizing transformation $f(\cdot)$ may be useful. The key function incorporates one of the standard choices, so that we may again take $f(\cdot)$ to be the identity. The link function is then $g(y) = \mathcal{K}^{-1}(y) = \sqrt{\zeta}\sin(y)$ with derivative $g'(y) = \sqrt{\zeta}\cos(y)$. The proportions p_1 and p_2 are estimated by $\tilde{p}_1 = (X_1 + 0.5)/(n_1 + 1)$, $\tilde{p}_2 = (X_2 + 0.5)/(n_2 + 1)$, and substituted into the formulae for Δ, p and ζ, to obtain estimated transformed effects $Y = \mathcal{K}(\tilde{\Delta})$.

If, alternatively, a linear model is assumed for correlation effect sizes $\rho = \Delta/\sqrt{\zeta}$, take $g(y) = \sin(y)$.

14.2.3 Log relative risk (two Poisson rates)

Consider a meta analysis of K large studies of a rare disease. The numbers of observed cases X_2 and X_1 can be modeled by Poisson random variables, and the relative risk (RR) is the ratio of Poisson rates $\rho = \mu_2/\mu_1$ of the two arms of the study. This basic model was considered in Section 9.1.2.

For each study, conditionally on the total number of responses $X_1 + X_2 = w$, the number of cases under treatment follows a binomial distribution,

$$X_2 \text{ given } X_1 + X_2 = w \sim B(w, p).$$

As in the previous case, various linearizing transformations $f(\cdot)$ are considered in the literature. One of the standard choices is the logarithm of the relative risk $\ln(\rho)$ related to the parameter p via linearizing transformation $\ln(\rho) = \ln(q^{-1} - 1) - \ln(p^{-1} - 1)$. For the meta-regression, this quantity is chosen as the response variable.

The appropriate key is

$$\mathcal{K}(p) = \arcsin(1 - 2q) - \arcsin(1 - 2p), \tag{14.4}$$

and the inverse function is $\mathcal{K}^{-1}(x) = (1 - \sin(\arcsin(1 - 2q) - x))/2$. The link function is thus

$$g(y) = f(\mathcal{K}^{-1}(y)) = \log(q^{-1} - 1) - \log\left(\frac{1 + \sin(C - y)}{1 - \sin(C - y)}\right), \tag{14.5}$$

where $C = \arcsin(1 - 2q)$. The proportion p is estimated by $\tilde{p} = (X_2 + 0.375)/(w + 0.75)$, and it is used to obtain estimated transformed effects $Y = \mathcal{K}(\tilde{p})$; a linear model is fitted for $g(Y) = \log\tilde{\rho} = \log((q^{-1} - 1)(\tilde{p}^{-1} - 1)^{-1})$. The inverse link function is

$$g^{-1}(y) = C - \arcsin\left(\frac{Re^{-y} - 1}{Re^{-y} + 1}\right)$$

where $R = q^{-1} - 1$ and the derivative is

$$g'(x) = \frac{2}{\cos(C - x)}.$$

14.3 Examples

This section presents two examples of meta-regression. First an example of meta-regression for standardized differences in means taken from Section 8.F.2 of Hedges and Olkin (1985) is considered. Then a meta-regression for log relative risk of tuberculosis originally reported by Colditz *et al.* (1994) is refitted.

14.3.1 Effect of open education on student creativity

Effect size estimates from $K = 10$ studies of the effects of open versus traditional education on student creativity are given in the Table 14.1. The covariate of interest is the grade level. Sample sizes for both modes of education are equal in each study, and the effect size is $\hat{d}_{\text{Cohen}} = (\bar{X}_2 - \bar{X}_1)/s_{\text{pooled}}$. The variances in the two arms are assumed to be equal (see Hedges and Olkin, 1985, pp. 185–187). Hedges and Olkin use the *vst* transformation with the key function $\mathcal{K}(d) = g^{-1}(d) = \sqrt{2}\sinh^{-1}(d/(2\sqrt{2}))$ before performing the standard linear model analysis, which tests for the difference in grades 1–3 (coded 1) versus grades 4–8 (coded 2). This corresponds to the regression model (14.2).

Hedges and Olkin found the effect of open education to decrease significantly in grades 4–8, with mean −0.327, and 95 % confidence interval [−0.484, −0.170]. The residual sum of squares (RSS) for their model is 31.285; and this model (which includes grade differences only) is rejected at the 0.01 level using a chi-squared distribution with $K - u = 8$ degrees of freedom. A linear regression may be a better model, providing an intercept of 0.5202 and slope of −0.1106 per grade, with RSS equal to 27.821 still with 8 degrees of freedom.

Table 14.1 Effect size estimates from 10 studies of the effects of open versus traditional education on student creativity, reproduced with minor changes from Table 4, Section 8.F.2 of Hedges and Olkin (1985). The values of $\hat{d}_k$ are in fact slightly corrected unbiased standardized mean differences, but the difference between these values and d_{Cohen} values is uniformly less than 0.02 across the table and is therefore ignored.

Study	Grade level	$n_2 = n_1$	$\hat{d}_k$	$\hat{\kappa}_k$
1	6	90	−0.581	−0.288
2	5	40	0.530	0.263
3	3	36	0.771	0.381
4	3	20	1.031	0.505
5	2	22	0.553	0.275
6	4	10	0.295	0.147
7	8	10	0.078	0.039
8	1	10	0.573	0.284
9	3	39	−0.176	−0.088
10	5	50	−0.232	−0.116

The other model (14.1) uses a normal GLM with link function $g(y) = 2\sqrt{2}\sinh(y/\sqrt{2})$. Note that $q = 1 - q = 1/2$ for each study. Both methods provide similar answers.

The GLM approach with the response variable $Y = \hat{\kappa} = g^{-1}(d)$ also shows a significant decrease in the effect of open education in grades 4–8, with mean −0.657 and 95 % confidence interval [−0.974, −0.339]. Residual deviance is exactly the same value of 31.285 we had before, and the lack-of-fit evidence is rather strong at 3.259. Using grade as a continuous predictor, the model equation is $d = g(Y) = 1.053 - 0.224\,\text{grade}$, with adjusted confidence intervals for the regression coefficients being [0.589, 1.516] for the intercept, and [−0.322, −0.125] for the slope. The residual deviance is 27.742 with 8 degrees of freedom, and the lack-of-fit evidence is 2.91. This is still not a perfect model. The plot of fitted versus observed values of $Y = g^{-1}(d)$, and the QQ plot of the residuals are shown in Figure 14.1. The QQ plot is much better than for the model with dichotomous grade level.

14.3.2 Vaccination for the prevention of tuberculosis

The data from $K = 13$ RCTs each comparing a group vaccinated by Bacillus Calmette-Guerin (BCG) vaccine for the prevention of tuberculosis against a nonvaccinated group, originally reported by Colditz *et al.* (1994), were already considered in Section 9.2.1 and reproduced in Table 9.1. It was suspected that the distance from the equator affected the efficacy of the vaccine, and therefore this covariate is to be investigated in the meta-regression. Latitude was centered by subtracting its mean (33.46). Since

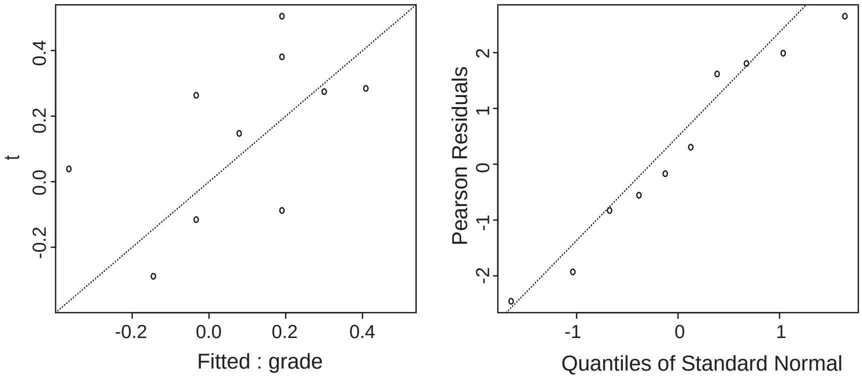


Figure 14.1 Plot of fitted versus observed values of $Y = g^{-1}(d)$, and the QQ plot of residuals for the GLM model of effects of open versus traditional education on creativity with grade as a continuous covariate.

only the distance from the equator and not the sign is of interest, the negative sign in study 9 carried out on the opposite side of the equator was dropped for the analysis.

14.3.2.1 Standard meta-regression with fixed effects

In the original study the log(RR) of disease in vaccinated group, defined as $\theta = \log \rho = \log(\mu_2/\mu_1)$ was the response variable. Here the index 2 corresponds to 'vaccinated' and 1 to 'nonvaccinated'. The inverse variances of log(RR) were used as weights in a standard weighted regression based on a normal approximation to log(RR). The answer was

$$\hat{\theta} = -0.635 - 0.029(x - 33.46),$$

where x is the distance from the equator in degrees latitude. As the distance from the equator increases the log(RR) decreases, corresponding to greater vaccine efficacy. To use correct tests and confidence intervals, the standard errors for the coefficients were divided by the MSE 1.672. The adjusted confidence intervals for intercept and slope are $[-0.722, -0.547]$ and $[-0.034, -0.024]$, respectively. The RR for the average distance from the equator observed in the trials is an exponent of the intercept estimate, which is 0.530. Similarly, the confidence interval for the RR is $[e^{-0.722}, e^{-0.547}] = [0.486, 0.578]$.

The plot of log (RR) versus the distance from the equator, and the QQ plot of the residuals are shown in Figure 14.2. The radii of circles on the left-hand plot correspond to the weights of trials. The model is driven by trials 6, 8 and 11. The QQ plot shows three outliers.

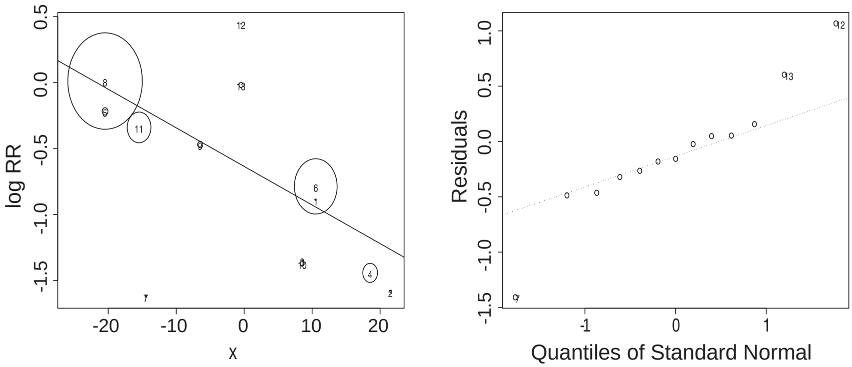


Figure 14.2 Plot of distance from the equator versus log (RR) with regression line log(RR) $= a + bx$, and the QQ plot of residuals. The radii of circles on the left-hand plot correspond to the weights of trials. The labels are trial numbers. The model is driven by trials 6, 8 and 11.

14.3.2.2 Meta-regression based on conditional standardized effects

The use of the Poisson approximation to the binomial is discussed in Section 18.4. Following a recommendation of Decker and Fitzgibbon (1991) it can be used only for small probabilities satisfying $p < 0.47/n^{0.31}$. For the BCG data this condition is satisfied for all trials except trial 6, in which the probabilities in both arms are too large. In trial 2 the proportion of disease in the not vaccinated arm $p = 0.1$ is only slightly higher than $0.47/n^{0.31} = 0.08$. The data-generating mechanism can thus be

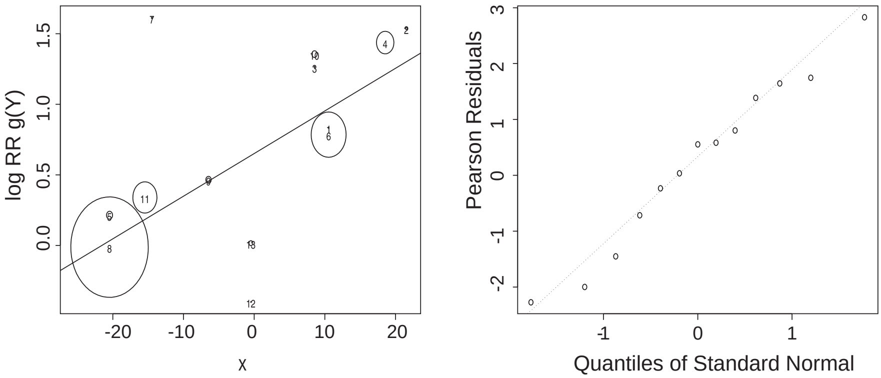


Figure 14.3 Plot of distance from the equator versus log RR $g(Y)$ with the linear fit $g(Y) = a + bX$ from GLM, and the QQ plot of residuals. The radii of circles on the first plot correspond to the inverse numbers of cases in trials. The labels are trial numbers.

approximated by the Poisson distribution, and the conditional key function (14.4) can be used. This is a more adequate approach since transformed standardized effects $\mathcal{K}(p)$ are assumed to be normally distributed, whereas the log (RRs) themselves are not. Let us fit a model for log (RR) in the nonvaccinated group using the generalized linear model (14.1) and the link function (14.5). Let w be the total number of cases in each trial, and X_2 and X_1 be the numbers in each subgroup (nonvaccinated versus vaccinated). Response variable Y is a vector of transformed standardized effects for conditional evidence (given total number w of cases in each trial) calculated as $Y = \arcsin(1 - 2q) - \arcsin(1 - 2\tilde{p}))$, with $\tilde{p} = (X_2 + 0.375)/(w + 0.75)$.

The fitted model is

$$g(Y) = \log \mathrm{RR} = 0.6513 + 0.0302(x - 33.46).$$

Note that the coefficients are very close to those from the standard meta-regression. The sign is opposite because in the original model the RR was defined as μ_2/μ_1, and in (14.4) it was defined as as μ_1/μ_2. The estimated dispersion parameter is $\sigma^2 = 2.618$. The standard errors need to be divided by $\sqrt{2.618}$. Adjusted confidence intervals for the regression coefficients are [0.561 to 0.742] and [0.0249 to 0.0355], respectively. The width of these confidence intervals hardly differs from the width of confidence intervals for the standard meta-regression.

The plot of log RR $g(Y)$ as a function of the distance from the equator, and the QQ plot of residuals are shown in Figure 14.3. The QQ plot is much better than for the previous model. Its superiority to the QQ plot from the standard meta-regression results from almost true normality of transformed standardized effects as opposed to dubious normality of log RR.

The null deviance is 164.2759 on 12 degrees of freedom, and the residual deviance is 28.8021 on 11 degrees of freedom. Thus the evidence for badness of fit is 2.53. This is certainly not a perfect model as can be seen from the spread on the plot.

15

Accounting for publication bias

A well-planned study may fail to generate the hoped-for amount of evidence. The reasons may include an insufficient sample size, a smaller than expected effect size, or imprecise and highly variable measurements of the influence of the treatment. By combining several low-powered studies, stronger evidence may be obtained. This is the idea that underlies meta-analysis. Publication bias is in some sense an inverse outcome. A result enters the published record with a claimed evidence that is exaggerated. This can be caused by a selection bias. If small studies are run repeatedly, one or a few of them may produce weak evidence. Because the published weak evidence is the maximal amount observed in repeated trials, this can happen even if the null hypothesis of a zero effect size is true.

If we know something about the selection mechanism, the published evidence can be corrected. The use of a *vst* that leads to approximate normality with fixed variances simplifies the necessary computations.

15.1 The downside of publishing

Data and model

- In a meta analysis one combines the available and comparable studies in order to obtain a more precise estimate of an effect. We assume that the estimated evidences of K studies have been published. Thus, K couples of study sizes and evidence values (n_k, T_k) for $k = 1, \ldots, K$ are at our disposal. These can be combined to give evidence

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

$$T_{\text{combined}} = \frac{\sqrt{n_1}\, T_1 + \cdots + \sqrt{n_K}\, T_K}{\sqrt{n_1 + \cdots + n_K}}, \tag{15.1}$$

for which the sum $n = n_1 + \cdots + n_K$ is the appropriate sample size. The combined effects found by a meta analysis often appear to overstate the evidence and to be biased in favor of the alternative. Such a bias could be due to a nonrandom selection of the studies, for example by favoring those studies that show a large effect. Publication bias is the name given to such a selection. Because the meta analyst only has access to published studies and because studies are only published if they show a significant effect, a selection bias is created.

- To model the publication bias, we suppose that the observed evidences satisfy

$$T_k \sim \mathcal{TN}(\sqrt{n_k}\,\kappa, 1, 1.645),$$

where $\mathcal{TN}$ denotes the truncated normal distribution with center $\sqrt{n_k}\,\kappa$, variance 1 and truncation point 1.645. Truncation at 1.645 means that all evidences smaller than 1.645 are absent. The truncated normal density is constructed from a normal density by setting the value of the density equal to zero to the the left of the cutoff point. The resulting curve is not a density, because it encloses an area of less than one. To make it into a density, one multiplies by the necessary constant.

This is the simplest possible explanation of the selection bias. To make the model more general, we could choose another truncation point or make the truncation point depend on the study. We could also make the effect size, expressed by the value of κ, depend on the study. But the positive side of such modifications – they render the model more realistic – have to be balanced with the negatives: they complicate the model's use and make its results less transparent.

Under the truncated normal model, some studies are absent from the published record and we take care of this by assuming knowledge about the mechanism for truncation. A better model is obtained by adding an additional feature, the number of missing or absent studies. The user of this model must specify the number of missing studies. By doing this, one gains control over the bias correction introduced in the meta analysis. Adjusting the cutoff point would serve a similar purpose.

Question

- If Equation (15.1) gives a biased account of the combined evidence, how can we correct for the bias?

Bias correction

- Based on the sample (n_i, T_i) $(i = 1, \ldots, K)$, the cutoff point and the number of missing studies, a statistical estimate of κ can be derived. The details of the calculation are explained in Chapter 26. Let $\hat{\kappa}_{\text{meta}}$ be this estimate. The bias-corrected evidence estimate is then simply

$$T_{\text{meta}} = \sqrt{n_1 + \cdots + n_K}\, \hat{\kappa}_{\text{meta}}.$$

Interpretation

- By construction, because the meta analysis takes the missing studies into account and corrects a bias in favor of the alternative, we have $T_{\text{meta}} \leq T_{\text{combined}}$. By increasing the number of missing studies from zero, a decrease in T_{meta} can be observed and we recommend to compute the value for several choices.

15.2 Examples

15.2.1 Environmental tobacco smoke

Tweedie *et al.* (1996) give an example of relative risk estimates based on 36 case-control studies. The disease these studies considered was lung cancer and the risk factor was environmental tobacco smoke (ETS). The data given in the paper are unadjusted risk ratios. The published values range from 0.74 (no risk, ETS decreases the occurrence of the disease) to 2.55 (large increase in risk for lung cancer due to ETS). In Chapter 7 the transformation to evidence of risk estimates has been discussed. There, risk is defined as the difference $p_1 - p_2$, where p_1 is the probability of the disease for the group with the risk factor activated and p_2 is the chance for those with the risk factor absent. The relative risk on the other hand is equal to the ratio $\text{RR} = p_1/p_2$.

In order to prepare a data set suitable for our purpose, we needed a way to get from one to the other. Assuming a value for p_2, this is easy and we find $p_1 - p_2 = p_2(p_1/p_2) - p_2 = p_2(\text{RR} - 1)$. From the published data set, the sample sizes are not known either. We made the assumption that each study used an equal number $n_1 = n_2 = n$ of cases and controls. We were then able to infer the value of n from the length of the confidence intervals for the relative risk given in Tweedie *et al.* (1996).

Let $R_i = p_2(\text{RR}_i - 1)$ and $N_i = n_i + n_i$ denote the values for the risk and the sample size in the ith study. The evidence is then – up to the small sample corrections, which are not important in this example – given by

$$T_i = \sqrt{N_i}\ \text{arcsin}\ \left(R_i/\sqrt{4 \times p_i(1 - p_i)}\right),$$

where $p_i = (p_{1,i} + p_2)/2 = p_2\,(\text{RR}_i + 1)/2$.

Conversely, we can infer the value of the relative risk RR from the evidence T and the sample size N by solving the following equation:

$$\frac{p_2\,(\text{RR} - 1)}{\sqrt{4\,p_2\,(\text{RR} + 1)/2(1 - p_2\,(\text{RR} + 1)/2)}} = \sin\,(T/\sqrt{N}\,).$$

If there were selection bias, we would expect some studies at the lower end of the evidence scale to be absent. A graphical inspection shows that the distribution of the 36 evidence values has a longer tail than the normal and its variance is smaller than one. This could be the result of variation in the effect size κ between the studies.

The combination of the 36 evidence values leads to $T_{\text{combined}} = 2.24$, which corresponds to a relative risk of 1.12. Assuming trunction with a known number of missing studies (censoring) and truncation point 1.645 leads to minimal corrections. When the number of missing studies is for example set to five, we obtain $T_{\text{meta}} = 2.19$, which corresponds to a relative risk of 1.119. In this example we conclude that there is weak evidence of an increase in lung cancer risk due to ETS.

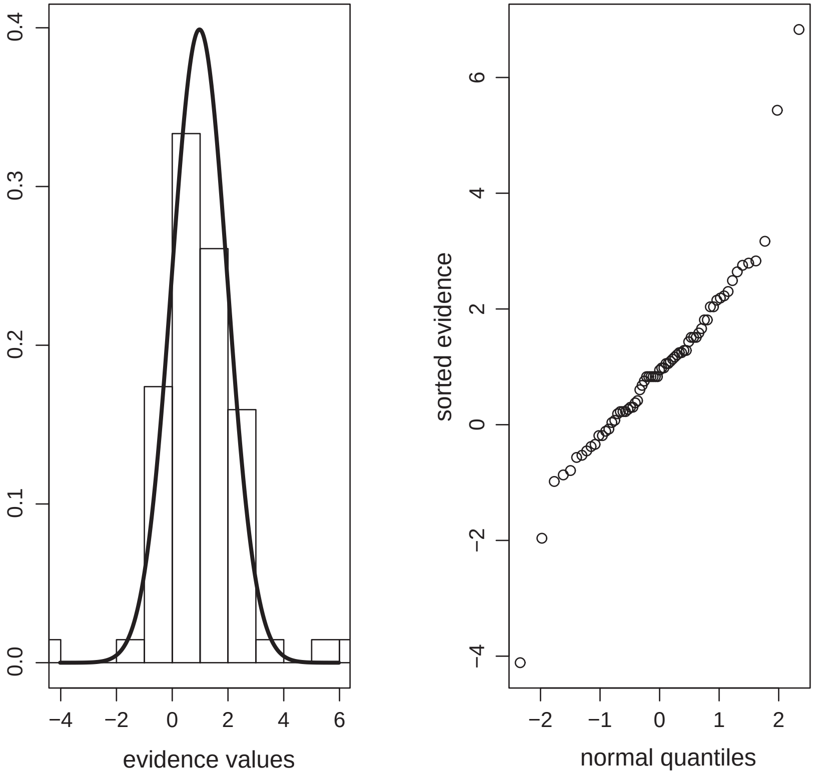


Figure 15.1 This plot shows the evidences obtained from 69 studies (Jané-Llopis *et al.* (2003)). In the paper, the standardized mean differences are tabulated. We assumed a sample size of 200 individuals in each study and converted the standardized mean differences to evidence by a simple rescaling. In the left-hand panel the normal density and an overlaid histogram of the evidence values is shown. In the right-hand panel the sorted evidence values are compared to normal quantiles.

15.2.2 Depression prevention programs

Jané-Llopis *et al.* (2003) report on a meta analysis of 69 studies on the effectiveness of depression prevention programs. Figure 15.1 shows the sorted evidence values versus normal quantiles. As pointed out in Jané-Llopis *et al.* (2003), the effects are very nearly normally distributed, with the exception of four studies that reported unusual findings. The plot reveals no publication bias due to suppressed studies near the lower end. Since the paper does not contain sample sizes, we had to assume values in order to apply the publication bias correction.

The combined evidence from the 69 studies equals 8.1 and speaks strongly in favor of a positive effect. Assuming five missing studies modifies this value only slightly downwards to 7.9.

The two least significant studies have evidence values of -4 and -2. Deleting these from the meta analysis results in an increased combined evidence of 9.0 instead of 8.1. Correcting for publication bias assuming two missing studies with a cutoff at evidence $= -1.5$, one obtains a corrected combined evidence of 8.3, which is surprisingly close to the original value.

We have used these studies as an example for publication bias even though the interest centers primarily on the differences between the study characteristics and the influence of these differences on the outcome.

Part II
The Theory

16

Calibrating evidence in a test

Many scientists regard the *p*-value as a measure of evidence against a null hypothesis, while others regard evidence better encapsulated in a confidence interval for an effect. Two advantages of the *p*-value are that only one number is specified, and it has a wide range of applicability. But the *p*-value also has numerous deficiencies which have been widely documented: see, for example, Schervish (1996) and Goodman (1998).

For us, the evidence for an effect lies in a test statistic, and to measure the evidence requires only a transformation to a simple calibration scale. On this scale the evidence always has a standard normal error in estimating its expected value. Thus only one number, the evidence for the effect, is reported, and it is always accompanied by a known error distribution which is familiar to all students of statistics. Interpretation of evidence is then more natural and easily communicated to others. And interpretation is possible under alternative hypotheses, extending the range of its usefulness.

This procedure further leads to confidence intervals for the effect, and facilitates combination of evidence for the same effect from different studies. Having a simple calibration scale allows for concentration on other important statistical issues of how to choose alternative hypotheses, and whether to allow for different or even random effects in combining evidence from different studies.

The price paid for calibrating evidence on such a simple scale is that one has to get there. This is done in theory by taking large enough samples so that the test statistic or estimator is approximately normal. But as many early statisticians pointed out, one can also get there with much smaller sample sizes by means of a variance stabilizing transformation. What this means in practice is that by applying a variance stabilizing transformation to the test statistic, one can often achieve approximate normality, with

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

standard deviation one, for much smaller sample sizes than required by central limit theorem approximations.

Typically evidence will be positive for a positive effect and negative for a negative effect, that is, an effect in the opposite direction. One must allow for both, especially when combining studies, so that conflicting results are allowed to cancel out. Because of the symmetry of the normal calibration scale, it suffices to define evidence for one direction.

16.1 Evidence for one-sided alternatives

Let X have the normal distribution with mean μ and variance σ^2, hereafter denoted $X \sim N(\mu, \sigma^2)$. In this chapter we assume μ is unknown and $\sigma = \sigma_0$ is known. One may be interested in either testing a hypotheses regarding μ, or in estimating μ, and these two problems are connected through the notion of statistical evidence.

For testing the null hypothesis $H_0 : \ \mu = \mu_0$ against the one-sided alternative $H_1 : \ \mu > \mu_0$ we want a measure of the evidence against H_0 in favor of H_1. Or, if we do not have enough information to assume a one-sided alternative (the usual case), then we want a measure of the evidence against H_0 in favor of the two-sided alternative $H_2 : \ \mu \neq \mu_0$. This latter problem is postponed until Section 17.4 because first we need to get the calibration scale right.

Given a random sample of observations $X_1, \ldots, X_n$ which are independent and each distributed $N(\mu, \sigma_0^2)$, the usual estimator of μ and also a test statistic is $S = \bar{X}_n = \sum_i X_i/n$. One rejects the null when S is large, because it is clear that large values of S favor H_1 over H_0. But what is the right calibration scale for the evidence in S against H_0?

When estimating μ the standard error of $\bar{X}_n$ is $\sigma_0/\sqrt{n}$, which decreases at the rate $1/\sqrt{n}$, so one must effectively quadruple the experimental effort to double the accuracy of the estimator of the unknown μ. We take it as axiomatic that evidence in favor of an alternative regarding μ must grow at the same rate. It will be convenient for what follows to define the *effect* by $\theta = \mu - \mu_0$ and the *standardized effect* by $\delta = \theta/\sigma_0$. In this case there is an obvious choice for estimating θ, namely $\hat{\theta}_n = \bar{X}_n - \mu_0$. It is evident that the null and alternative hypotheses can be restated in terms of θ or δ; for example $H_0 : \ \theta = 0$ against $H_1 : \ \theta > 0$. This simple model is called the *prototypical model*.

We now define *one-sided evidence* against the null in favor of the positive alternative as any monotonically increasing transformation $T = T(S)$ of the test statistic S for which $T \sim N(\mathrm{E}[T], 1)$; that is, for which T is on the *unit normal* scale for all values of the parameters. A consequence of this definition is that evidence always has a normal distribution with fixed standard deviation of 1, facilitating comparisons between and combinations of evidence. Another is that the evidence T is closely identified with its expectation, in this case its mean $\tau = \mathrm{E}[T] = \sqrt{n}\,\theta/\sigma_0$. The standard error of T in estimating τ is 1.

In our simple model we can take $T = \sqrt{n}\,(\bar{X}_n - \mu_0)/\sigma_0$, which is sometimes called the Z-test statistic. Clearly $T \sim N(\tau, 1)$ for all μ. As a simple example, fix

$\mu_0 = 5$ and $\sigma_0 = 5$. Then for $n = 4$ and $\bar{X}_4 = 10$, the evidence against the null in favor of the positive alternative is $T = 2$, with standard error 1 when considered an estimator of the unknown τ. For $n = 36$ and $\bar{X}_{36} = 10$, the evidence against the null is $T = 6$, also with standard error 1.

The one-sided alternative $\mu < \mu_0$ can be treated symmetrically, by replacing T by $-T$. That is, negative one-sided evidence for $\mu > \mu_0$ is positive one-sided evidence for $\mu < \mu_0$. Evidence for two-sided alternatives is defined in Section 17.4.

16.1.1 Desirable properties of one-sided evidence

The reader will no doubt question the generality of the above definition, for once the standard deviation σ_0 of the observations X_i is unknown, or the distribution non-normal, the test statistic S will have a distribution which is non-normal with variance depending on unknown parameters. However, in many practical examples T can be chosen to stabilize the variance to 1, and simultaneously yield approximate normality. We list below four desirable properties E_1 to E_4 for a measure of evidence, which in practice are attained only to a certain, but usually sufficient, degree to measure the evidence against the null hypothesis and for an alternative.

Let θ be an unknown effect for which it is desired to test $\theta = 0$ against $\theta > 0$, and let S be a test statistic which rejects H_0 for large values of S. We want a measure of one-sided evidence T to satisfy

- E_1, the one-sided evidence T is a monotonically increasing function of S;
- E_2, the distribution of T is normal for all values of the unknown parameters;
- E_3, the variance $\text{Var}[T] = 1$ for all values of the unknown parameters; and
- E_4, the expected evidence $\tau = \tau(\theta) = \text{E}_\theta[T]$ is monotonically increasing in θ from $\tau(0) = 0$.

In the simple example of a normal model with known variance all of the above properties hold exactly for evidence defined by the Z-test statistic; that is, estimated standardized effect. In general, properties E_2 to E_4 will hold only approximately, but to a surprising degree, even for small sample sizes.

16.1.2 Connection of evidence to *p*-values

The p-value for an observed $S = s$ is computed by $p = P_0\{S \geq s\}$, where P_0 is the null distribution of S. Further, if $T = T(S)$ satisfies properties E_1 to E_3, then the p-value can also be computed from the observed value of $T = t$ by $p = P_0\{T \geq t\} = \Phi(-t)$, so $t = t(p) = \Phi^{-1}(1 - p)$. Table 16.1 contains some values of $t(p)$ for comparison with p. As a significant promoter of p-values, Fisher (1926) originally suggested the level 0.05, saying

Table 16.1 Selected values of p, $t(p) = \Phi^{-1}(1 - p)$ and the ratio $t(p)/t(0.05)$. The second row is on the probit scale while the third row uses 0.05 as a reference point. The traditional markers 0.05, 0.01 and 0.0005 represent evidence in proportions $1 : \sqrt{2} : 2$.

p	0.0005	0.001	0.01	0.02	0.025	0.05	0.1	0.1587
$t(p)$	3.291	3.090	2.326	2.054	1.960	1.645	1.276	1.000
$t(p)/t(0.05)$	2.000	1.879	1.414	1.248	1.192	1.000	0.779	0.608

> Personally this writer prefers to set a *low standard* of significance at the 5 per cent. point, and ignore entirely all results which fail to reach this level. [Our emphasis.]

We somewhat arbitrarily describe values of T near 1.645 as *weak* evidence against the null. Values of T which are twice as large we call *moderate* evidence, and values which are three times as large as *strong* evidence. Thus our definition of weak evidence follows Fisher's low standard when the null is true, but we are otherwise measuring evidence against the null on a different calibration scale, one which allows for interpretation whether or not the null hypothesis holds.

Now the observed $T = t$ is a monotonic function of the p-value; but the salient difference between t and p is that under alternatives the p-value distribution is highly skewed and changing with sample size, making interpretation and combinations of evidence difficult, as explained in Section 16.2.

16.1.3 Why the *p*-value is hard to understand

16.1.3.1 The *p*-value is a conditional probability

The definition of the p-value requires four ingredients: first, a null hypothesis about the state of nature; second, a test statistic S which orders the outcomes of an experiment, with the larger the value of S, the more evidence against the null hypothesis; third, the probability distribution of the test statistic when the null hypothesis holds; and fourth, an observed value of $S = s$ from the experiment. The p-value of the outcome $S = s$ is then defined to be the probability under the null hypothesis that $S \geq s$. The evidence is in the test statistic S and the p-value is a measure of 'surprise', with smaller values of the p-value raising the question of whether the null hypothesis could in fact be true. Note that the word 'evidence' is used in the everyday sense of the word.

The reason Fisher promoted the p-value is that he found it useful for discarding unremarkable experimental results from those which might be worth further consideration. Based on his experience with many experiments, he gave guidelines for what might be considered significant p-values. But he did not intend that 0.05 should become a standard for publication, or that p-values should be compared or used to predict future experimental results.

However, given its widespread adoption it was perhaps inevitable that not only statisticians but scientists in general would start trying to interpret it from frequentist or Bayesian points of view. In particular, a scientist might ask whether one can expect to obtain a similar *p*-value in an identical replication of the experiment; and if not, why not (see the end of Section 16.2.2). Or, because *p*-values are often interpreted naively as the probability of the null hypothesis, given the data, one might ask whether these concepts have anything to do with each other (see Section 16.4).

The above definition of the *p*-value is a *conditional* probability, computed given an event $S = s$ in a specific experiment and therefore applicable only for that experiment. It does not have any further interpretation. If one wants to interpret it unconditionally, that is, from outside the particular experiment which led to it, one needs to define it differently.

16.1.3.2 The unconditional, or random *p*-value

Let S_0 be independent of S and have the null distribution of S. This S_0 represents the outcome of an independent repetition of the experiment, in which conditions are identical to those of the original experiment, and in which the null hypothesis holds. Then given $S = s$ in the experiment just conducted, define the *p*-value by $P(S_0 \geq s)$. This yields the same conditional *p*-value as above, because S and S_0 have the same distribution under the null hypothesis.

Now define the *random p-value* by $PV = P(S_0 \geq S) = 1 - F_0(S)$, where for simplicity of presentation we assume the null cumulative distribution function F_0 (the cdf of S_0), is continuous. The cdf of PV, for $0 < p < 1$, is

$$\begin{aligned} F_{PV}(p) &= P\{PV \leq p\} = P\{1 - F_0(S) \leq p\} \\ &= P\{F_0(S) \geq 1 - p\} = P\{S \geq F_0^{-1}(1 - p)\} \\ &= 1 - F_1(F_0^{-1}(1 - p)), \end{aligned} \tag{16.1}$$

where F_1 is the cdf of S. Note that this definition does not require F_1, the distribution of the original test statistic S, to be the same as the null distribution F_0. When it does ($F_1 = F_0$), it follows from (16.1) that $F_{PV}(p) = p$ for $0 < p < 1$, so the random PV has the continuous uniform distribution on the interval [0,1]. When F_1 differs from the null distribution, the random PV often takes on a very different distribution.

16.1.3.3 Random *p*-value for the prototypical model

Let $X \sim N(\mu, 1)$, with the hypotheses of interest being $\mu = 0$ and $\mu > 0$. Given $X = x$, the ordinary (conditional) *p*-value is $P(X_0 \geq x) = 1 - \Phi(x)$, where X_0 has the null distribution of X and is independent of it. The cdf of X depends on μ and is $F_\mu(x) = P\{X \leq x\} = P\{X - \mu \leq x - \mu\} = \Phi(x - \mu)$. The (unconditional) random *p*-value based on X is $PV(X) = 1 - \Phi(X)$. Substituting these results in (16.1), one obtains its cdf $F_{PV}(p) = 1 - F_\mu(t(p)) = \Phi(t(p) - \mu)$, where $t(p) = \Phi^{-1}(1 - p)$ is the probit transformation discussed in Section 16.1.2. This cdf will be useful in deriving properties of PV in the next section.

16.2 Random *p*-value behavior

This section explains in part why a new calibration scale for evidence in the p-value is desirable. For simplicity of presentation, let $\mu_0 = 0$ and $\sigma_0 = 1$ in the prototypical model. Thus $X \sim N(\mu, 1)$, the hypotheses of interest are $\mu = 0$ and $\mu > 0$ and the *random p-value* based on X is $PV(X) = \Phi(-X) = 1 - \Phi(X)$. The presentation is for $n = 1$ observation, but results for any n can be obtained by replacing X by $\sqrt{n}\bar{X}_n$ and μ by $\sqrt{n}\mu$.

16.2.1 Properties of the random *p*-value distribution

Let $z_q = \Phi^{-1}(q)$ denote the qth quantile of the standard normal distribution; that is, $q = P(Z \leq z_q)$, and define for each x the standard normal density by $\varphi(x) = \exp\{-x^2/2\}/\sqrt{2\pi}$. Then for the prototypical model the random p-value PV has the following properties:

- P_1. *The qth quantile of the distribution of $PV(X)$ is $p_q = p_q(\mu) = \Phi(z_q - \mu)$.* The notation $p_q = p_q(\mu)$ emphasizes that the qth quantile depends on μ. It may also be expressed in terms of the power of the Neyman–Pearson level-$(1 - q)$ test, namely $P_\mu(X > z_q) = 1 - \Phi(z_q - \mu) = 1 - p_q(\mu)$. With this formula one immediately sees that as μ increases without bound, the qth quantile $p_q(\mu)$ approaches 0 and the power approaches 1.

- P_2. *The expected value of the random $PV(X)$ is* $\mathrm{E}_\mu[PV(X)] = \Phi(-\mu/\sqrt{2})$. The reason for stating this formula is that it is common to describe a random variable in terms of its mean and standard deviation. This we have already done for the transformed p-value $t(PV)$, where we found the mean to be μ and the standard deviation 1. But these are good summary measures only when the distribution is symmetric or nearly so. The p-value distribution under alternatives is highly skewed, so the expected p-value is not a representative measure of its distribution. In fact, it follows easily from properties P_1 and P_2 that:

- P_3. *The expected p-value equals the qth quantile of its distribution, where q is given by $q = \Phi(\mu(\sqrt{2} - 1)/\sqrt{2})$.* For example, when $\mu = 1$, the expected p-value equals the $q = 0.61$ quantile of its distribution, and when $\mu = 3$ it equals the $q = 0.81$ quantile. Thus the expected p-value is totally unreliable for representing the p-value distribution under alternatives. The expected p-value was studied by Dempster and Schatzoff (1965) and more recently by Hung *et al.* (1997) and Sackowitz and Samuel-Cahn (1999); the latter authors also consider quantiles of the p-value distribution under alternatives and give an application of their findings. Median p-values are investigated by Bhattacharya and Habtzghi (2002).

- P_4. *The density of the p-value is* $f_{PV}(p) = \varphi(t(p) - \mu)/\varphi(t(p)),\ 0 < p < 1$, where $t(p) = \Phi^{-1}(1 - p)$ for all $0 < p < 1$. This formula is derived in Donahue (1999). Plotting $f_{PV}(p)$ against p for any $\mu > 0$ reveals it to be concave, monotonically decreasing and skewed to the right. Moreover, the plots change shape with μ, so it is difficult to make comparisons between different p-values under alternatives.

16.2.2 Important consequences for interpreting *p*-values

What are the implications of the above results P_1 to P_4 for interpreting the evidence in the p-value? A good way to grasp the implications is to consider some quantiles of the corresponding distributions. The p-value and corresponding probit-value $X = t(p)$ are related by a monotonically *decreasing* function so the qth quantile of one distribution transforms into the $(1 - q)$th quantile of the other. For example, the $q = 0.5$ quantile or median of X is μ, and this corresponds to the median p-value $\Phi(-\mu)$. When $\mu = 1.645$, the median p-value is $\Phi(-1.645) = 0.05$. The reader is asked to study the plot of $t(p)$ versus p in Figure 16.1, and find the corresponding quantiles of these two distributions.

Next take the $q = 0.75$ quantile of the evidence distribution which is $\mu + 0.6745$; it transforms into the $(1 - q) = 0.25$ quantile of the p-value distribution, which is $\Phi(-\mu - 0.6745)$. When $\mu = 1.645$ this quantile is $\Phi(-2.32) = 0.01$. Similarly the reader can check that in this example the 0.25 quantile $\mu - 0.6745$ of the evidence distribution transforms into $\Phi(-0.971) = 0.166$. Thus when $\mu = 1.645$, the central 50 % of the evidence distribution (shown shaded in Figure 16.1) corresponds to the 50 % of the p-value distribution lying between 0.010 and 0.166. The fact that the latter interval is not centered on the median p-value of 0.05 reflects the asymmetry of this distribution under alternatives.

Now let μ be unknown. Having observed $X = x$, a 50 % confidence interval for μ is of the form $[x - 0.6745, x + 0.6745]$. This interval transforms into a 50 % confidence interval $[\Phi(-x - 0.6745), \Phi(-x + 0.6745)]$ for the corresponding median p-value $\Phi(-\mu)$. For example, when $X = 1.645$ the observed p-value is 0.05, but a 50 % confidence interval for the median p-value is $[0.010, 0.166]$. The reader can similarly find intervals with different levels of confidence, but the message is clear: simply stating that the p-value is 0.05 gives the wrong impression that one is close to the mark.

An important result follows from property P_2. If one has conducted an experiment and obtained a p-value of 0.05, then the estimate of μ is 1.645 and hence the maximum likelihood estimate of the expected p-value is $\Phi(-1.645/\sqrt{2}) = 0.122$. Thus in a repetition of the experiment, the researcher can expect a p-value of 0.122. Similar findings are reported by Goodman (1992).

16.3 Publication bias

It is well known that the established practice of requiring experimental results to contradict a null hypothesis of no effect at level 0.05 introduces certain anomalies.

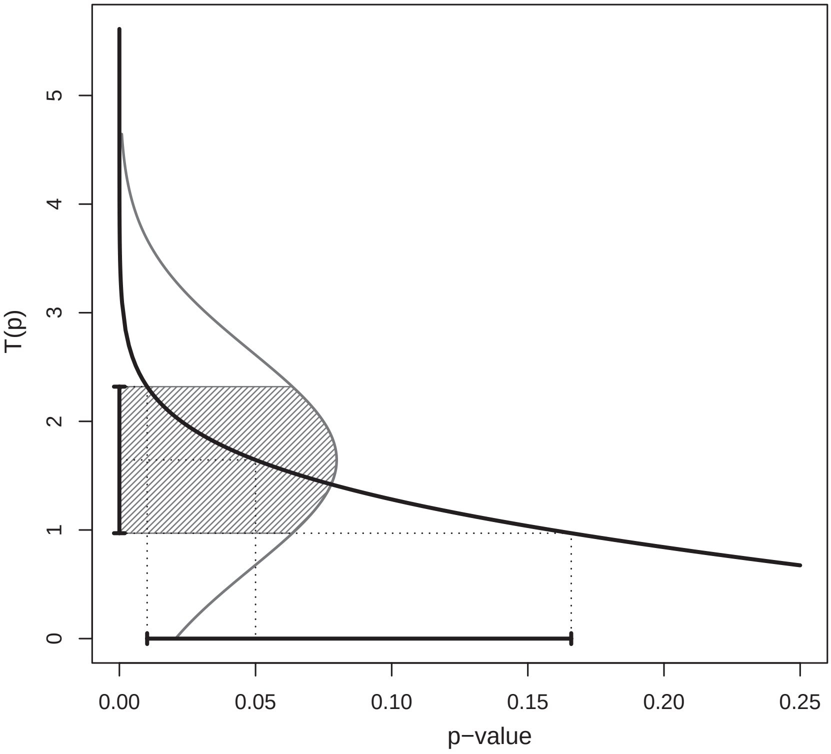


Figure 16.1 The curve shows the evidence $t(p)$ for each p-value. The p-value of 0.05 and its transformed value of 1.645 are highlighted. Around those values a 50 % confidence interval is drawn, both for the p-value itself and for the transformed p-value. Also indicated is how the interval boundaries are linked to each other. On the transformed scale, the random variation is the same, no matter the value of the transformed p-value. This is indicated by the normal density and the shaded area covering 50 % of the area.

The scientist who obtains a p-value of 0.049 may succeed in publishing the result, while the one who obtains 0.051, which is not publishable, knows there is just about as much evidence against the null in his or her data as that in the 0.049 result. The very fact of publication introduces a bias towards the alternative: a published p-value is *conditional* on its being less than a threshhold. Of course there are other factors which are more important in publishing than the size of the p-value, but here we only examine this one.

Assume evidence $T_n = \sqrt{n}\bar{X}_n$ for testing $\mu = 0$ against $\mu > 0$ in the prototypical model. Let $A_n = \{PV_n \leq 0.05\} = \{T_n \geq 1.645\}$ be the event that the evidence in the p-value is significant at level 0.05. The *conditional evidence* in the random p-value, given that it is significant, is defined to be $U_n = T_n|A_n$, where $P_\mu(A_n) = 1 - \Phi(c_{\sqrt{n}\,\mu})$, with $c_{\sqrt{n}\,\mu} = 1.645 - \sqrt{n}\,\mu$. It is clear that the distributions of T_n and

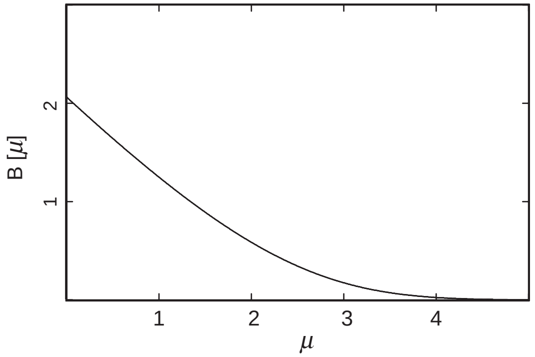


Figure 16.2 Plot of the publication bias function $B(\mu)$ versus μ. When $\alpha = 0.05$ the bias at $\mu = z_{0.95} = 1.645$ is a surprisingly large $B(1.645) = 0.8$.

U_n only depend on n and μ through $\sqrt{n}\,\mu$ so hereafter we only consider the case of of $n = 1$ and write $T = T_1$ and $U = U_1$. One can substitute $\sqrt{n}\,\mu$ for μ to recover the general case when desired.

If one restricts attention to p-values which are significant at level 0.05, because they are the ones available in the literature, then one exaggerates the evidence in such p-values by considering only U rather than T, effectively ignoring all nonsignificant results. The difference in means $B(\mu)$=E$_\mu[U]$−E$_\mu[T]$ will be called the *publication bias* at μ, for $\mu \geq 0$. It is shown in Chapter 26 that $B(\mu)=\varphi(c_\mu)/\{1-\Phi(c_\mu)\}$. A plot of this function is shown in Figure 16.2.

For small μ the publication bias is considerable: the average overstatement of evidence is more than 2 units on the probit scale. Of course, it is very unlikely (probability near 0.05) that when μ is small the p-value will be significant. Of more concern is that when $\mu = 1.645$, say, which is an effect of some interest, that the publication bias is 0.8. Estimation of publication bias and correction for it is the topic of Chapter 26.

16.4 Comparison with a Bayesian calibration

The p-value is often confused with the probability that the null hypothesis is true, given the data. For many frequentists, this confusion can best be resolved by education: in their view, the concepts of p-value and posterior probability are simply incommensurable. However, a considerable amount of research has gone into making such comparisons; see, e.g. Casella and Berger (1987), Berger and Sellke (1987), Berger *et al.* (1997), Selke *et al.* (2001), Hubbard and Bayarri (2003) and Berger (2003) and the discussions following them. It is therefore of interest to compare the calibration scale proposed here with a recent one by Selke *et al.* (2001).

These authors assume the p-value has the uniform density f_0 on [0,1] under H_0 and then consider alternative distributions $f_1(p)$ for the p-value under H_1. The likelihood ratio (or Bayes factor) for H_0 to H_1 is then $L(p) = f_0(p)/f_1(p)$, and, assuming a positive prior probability π_0 on H_0, the posterior probability of H_0, given p is $P(H_0|p) = \{1 + (1 - \pi_0)/\pi_0 L(p)\}^{-1}$. It is evident that for any $f_1(p)$ which is positive over [0,1] any desired value of $P(H_0|x)$ can be obtained by choice of prior probability π_0. These authors choose a 'default' value $\pi_0 = 0.5$ and show that (under some conditions) the likelihood ratio $L(p)$ is bounded below by $B(p) = -ep \ln p$ for $p < 1/e$ and 1 otherwise. This leads to a lower bound $\alpha(p) = \{1 + 1/B(p)\}^{-1}$ on $P(H_0|p)$. For example, when $p = 0.05$, $\alpha(p) = 0.289$. This leads them to conclude that the p-value overstates the evidence; it certainly does so if one uses $P(H_0|p)$ as a measure of evidence.

However, our thesis is that the p-value measures surprise, not evidence, so it is of interest to place the $\alpha(p)$ bound on the probit scale. A plot of $T(\alpha(p))$ against $T(p) = \Phi^{-1}(1 - p)$ is shown in Figure 16.3, for comparison with our calibration of the p-value. This graph shows that $\alpha(p)$ will typically underestimate the evidence in the p-value by 1 unit, the standard deviation of $T(PV_n)$, at least for significant p-values $p < 0.05$. In general $P(H_0|p)$ will exceed $\alpha(p)$, so a person using the smallness of

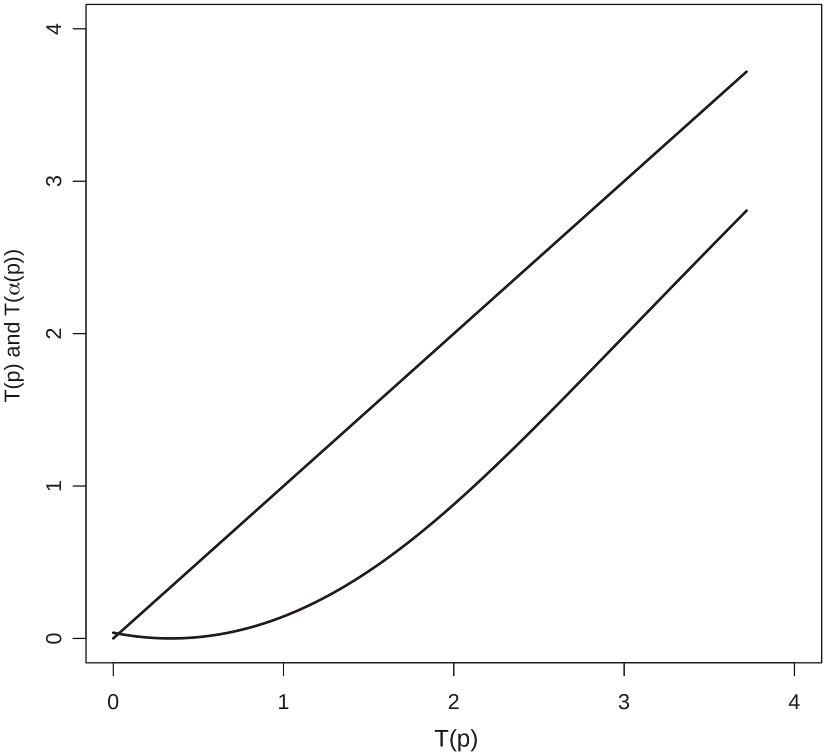


Figure 16.3 Plot of $T(\alpha(p))$ against $T(p)$ for $0.0001 < p < 0.5$. The region of interest is $T(p) \geq 1.645$, corresponding to $p \leq 0.05$.

$P(H_0|p)$ as a measure of evidence against the null will underestimate the evidence in the p-value by even more than 1 unit.

The difference between the frequentist approach to testing and the Bayesian approach espoused by Selke *et al.* (2001) is that these authors treat the hypotheses symmetrically. Contrast this with the frequentist approach which chooses one hypothesis to be the null so that the burden of proof is on the alternative, that is, because by definition Type I error is more important than Type II error.

We have considered the usual situation where the Type I error (making a false claim of an effect) is more important than the Type II error (not detecting an effect). But there are other problems where a large enough ‘effect’, positive or negative, of some proposed treatment is deleterious, and one will not adopt the treatment unless it proves otherwise. In this case appropriate hypotheses are null $H_0 : \; |\mu| \geq \mu_0$ and alternative $H_1 : \; |\mu| < \mu_0$. An inability to take into account the ramifications of hypothesis choice can lead to confusion and major mistakes (see Hoenig and Heisey 2001). We agree with Hoenig and Heisey that writers of modern textbooks would do well to emphasize the importance of choosing hypotheses carefully, and add that Neyman (1950) devoted four pages of his elementary text explaining how to choose appropriate hypotheses. The choice of hypotheses is determined by context, and they can rarely be interchanged in practice.

16.5 Summary

The p-value has been around for a long time because it has proven to be a simple and useful tool for filtering out very weak experimental results. But scientists and statisticians in particular want more from a measure of evidence. They want to be able to compare evidence from different experiments, and combine evidence from experiments testing for the same effect. When faced with a number of ‘significant’ results, each of which casts some doubt on the null hypothesis, it is natural to want to combine these results, *and to do so under an alternative hypothesis*.

While the random p-value is simple to interpret under the null hypothesis, under alternatives its distributions are highly skewed, making comparisons and combinations of results complicated. Nevertheless, we have learned a few things by looking at the p-value under alternatives for the prototypical model. Perhaps the most interesting one is that given a p-value of 0.05, the estimated expected p-value in an identical replication of the experiment is 0.12.

By transforming the random p-value onto the probit scale, one obtains a measure of evidence whose mean grows linearly with the effect and linearly with the square root of the sample size. On this scale a ‘highly significant’ p-value of 0.01 represents about 40 % more evidence than a p-value of 0.05. Further, one must face the fact that evidence contains random error, and on this proposed calibration scale it is always 1 unit, regardless of sample size or effect size. Often p-values are interpreted too precisely, perhaps because they are calculated to two or more decimal places. But the evidence in a conditional p-value of 0.05 would better be reported as evidence 1.645, with a standard error of 1.

We compared the evidence in a Bayesian calibration of the *p*-value and found that for all practical purposes, the posterior probability of the null, given the *p*-value, contains about one standard error less evidence than the *p*-value.

If one accepts the above proposal as potentially useful for thinking about evidence, the main remaining question is: how general are the above results? In many simple applications of statistics, variance stabilizing transformations which are already available will allow calibration on this scale, as we will demonstrate in the coming chapters.

17

The basics of variance stabilizing transformations

In this chapter we first review the simplest method for variance stabilization, standardization of the test statistic. Then we outline a general method for obtaining a *variance stabilizing transformation*, or *vst*, for short, and explain how we expect to benefit from it in finding evidence and confidence intervals. Then we will illustrate the theory with a *vst* for the sample mean estimator of the Poisson mean. Finally, we give an important example where a *vst* is desired: obtaining two-sided evidence from one-sided evidence on the probit calibration scale.

17.1 Standardizing the sample mean

Given the test statistic $\bar{X}_n$ based on a sample of n observations from a distribution with mean μ and variance σ^2, it is common, especially when n is large, to 'standardize' $\bar{X}_n$ by subtracting its mean, and dividing by its standard deviation to obtain a Z-statistic $Z_n = \sqrt{n}\,(\bar{X}_n - \mu)/\sigma$. This has three effects: firstly, it results in a variable centered at 0 ($\mathrm{E}[Z_n] = 0$) for all μ, σ; secondly, the variance is stabilized at 1 ($\mathrm{Var}[Z_n] = 1$ for all μ, σ); and thirdly, the distribution of Z_n is approximately standard normal, by virtue of the central limit theorem. Thus while $\bar{X}_n$ has variance σ^2/n taking on all positive values, Z_n has variance 1 for all values of the parameters. The transformation from $\bar{X}_n$ to Z_n thus 'stabilizes the variance'. The benefits of this transformation are clear, but it requires knowledge of the usually unknown parameter σ^2.

If σ^2 were replaced by the sample variance s^2 in the transformation, the resulting $Y_n = \sqrt{n}\,(\bar{X}_n - \mu)/s_n$ would not have a stable variance in the sense of being constant

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

for all μ, σ, but it may be close enough to 1 for practical sample sizes n and a corresponding range of μ, σ so that for the sake of inference about μ we may be able to act as though it were so.

Definition 17.1 *Any sequence of random variables* $\{Y_n\}$ *will be said to be* variance stabilized (to 1) *if* $\mathrm{Var}[Y_n] = 1 + c_n$*, where* $nc_n \to 0$ *as* $n \to \infty$*. This is sometimes written* $\mathrm{Var}[Y_n] = 1 + o(n^{-1})$*. The constants* $\{c_n\}$ *may well depend on model parameters, and the convergence to 0 is not necessarily uniform in the parameters. In any case we write* $\mathrm{Var}[Y_n] \doteq 1$.

In this book variance stabilization is about choosing transformations $h_n(S_n)$ of statistics $\{S_n\}$ that achieve the goal $\mathrm{Var}[h_n(S_n)] \doteq 1$. As indicated, h_n can depend on the known sample size parameter n. Note that $h_n(S_n)$ is itself a statistic.

In most cases the degree of approximation will be checked by simulations, even when a limit theorem exists that gives the rate of convergence of the variance of transformed variable to the target 1 as n increases without bound. We are mainly interested here in small and moderate sample sizes, and often drop the subscript n on h_n when it is clearly understood.

17.2 Variance stabilizing transformations

17.2.1 Background material

Let X denote a random variable with variance $\mathrm{Var}[X]$. Suppose $Y = h(X)$, where h is now any smooth function with at least two derivatives. Then the following expansions may be helpful when $\mathrm{Var}[X]$ is small:

$$\mathrm{E}[Y] = h(\mathrm{E}[X]) + \frac{h''(\mathrm{E}[X])}{2}\mathrm{Var}[X] + R_1; \tag{17.1}$$

$$\mathrm{Var}[Y] = \{h'(\mathrm{E}[X])\}^2\mathrm{Var}[X] + R_2. \tag{17.2}$$

Here R_1 and R_2 are remainder terms when the earlier terms on the right are used as approximations to the mean and variance of Y; these remainder terms will typically be of smaller magnitude than the earlier terms. These approximations will be used repeatedly throughout this book and can be found in Johnson *et al.* (1993, p. 54) or Bickel and Doksum (1990, p. 32); the latter reference also contains material on the error of approximation when the random variable X is a sample mean.

In our applications $X = S_n$ is a test statistic based on n observations and $h = h_n$ is chosen so that the transformed test statistic $Y = h_n(S_n)$ satisfies $\mathrm{Var}[h_n(S_n)] \doteq 1$. A first approximation to the expected value is then given by the first term in (17.1), and we write $\mathrm{E}[h_n(S_n)] \doteq h_n(\mathrm{E}[S_n])$ for this approximation. It is typically growing at the rate $\sqrt{n}$, while the bias term $h_n''(\mathrm{E}[S_n])\mathrm{Var}[S_n]/2$ is of smaller order, usually $1/\sqrt{n}$, and depending on unknown parameters. The remainders R_1 and R_2 are typically of order $n^{-3/2}$, also depending on unknown parameters.

Johnson *et al.* (1993, p. 54) or Bickel and Doksum (1990, p. 32) also point out a simple method for finding a function h so as to stabilize the variance, provided one can first write $\mathrm{Var}[X] = g(\mathrm{E}[X])$ for a known function g. One defines h as any

$$h(x) = \int^x [g(t)]^{-1/2} \mathrm{d}t, \tag{17.3}$$

provided the indefinite integral exists. Thus h is defined up to an additive constant. It follows that $\{h'(\mathrm{E}[X])\}^2 = \{g(x)\}^{-1} = \{\mathrm{Var}[X]\}^{-1}$, and then, by (17.2), we can expect that $\mathrm{Var}[Y] \doteq 1$. Thus, in principle, variance stabilization is easy, but in practice this method of finding h may not be fruitful, because the resulting h may depend on unknown parameters.

17.2.2 The Key Inferential Function

Let us call our parameter of interest θ. If S_n is a test statistic for $\theta = \theta_0$ versus $\theta > \theta_0$ for which large values of S_n lead to rejection (S_n could be an estimator $\hat{\theta}_n$), and h_n is a *vst* of S_n obtained from (17.3), we typically find that $h_n(S_n)$ has a variance near 1 for a desired range of values of θ, so it satisfies property E_3 of Section 16.1.1 of a measure of evidence. Further, it is often the case that h_n can be chosen to be monotonically increasing in its argument, so property E_1 is satisfied. In other words, $h_n(S_n)$ is still a test statistic for testing $\theta = \theta_0$ versus $\theta > \theta_0$. If it turns out that h_n depends on θ or unknown nuisance parameters, one can try substituting estimates for these parameters to see whether a measure of evidence can be obtained by modification of h_n. Hereafter we assume these hurdles have been overcome.

Property E_4 requires that the mean $\mathrm{E}[h_n(S_n)]$ be monotonically increasing in θ from 0 at θ_0. In many applications $\mathrm{E}[h_n(S_n)]$ is of the form $\sqrt{n}\, K(\theta)$ for n, θ of interest and K a known monotonically increasing function of θ. By subtracting the known constant $\sqrt{n}\, K(\theta_0)$ from h_n, we can ensure that $T_n = h_n(S_n) - \sqrt{n}\, K(\theta_0)$ will have a mean $\tau = \mathrm{E}[T_n]$ that satisfies E_4 as well as inheriting the properties E_1, E_3 from $h_n(S_n)$, because h_n is defined only up to an additive constant. Finally, we need to check that T_n also satisfies condition E_2, approximate normality for n, θ of interest. Having T_n approximately $N(\tau, 1)$ with $\tau = \sqrt{n}\, K(\theta)$ is highly desirable, because then T_n has a very well-known distribution and is an unbiased estimator of its mean τ, with standard error 1. In the text to follow we often write $T_n \sim N(\tau, 1)$, even though the distribution of T_n is only approximately normal.

Definition 17.2 *Given a statistical model and a measure of evidence T_n that satisfies properties E_1–E_4 of Section 16.1.1. Supposing further that its expected evidence $\tau = \mathrm{E}[T_n] \doteq \sqrt{n}\, \mathcal{K}(\theta)$, we call $\mathcal{K}$ the* Key Inferential Function *or simply the* Key *for this statistical model.*

The *Key Inferential Function leads to the solution of many routine problems:*

- K_1 Choosing the sample size n. *For testing $\theta = \theta_0$ against $\theta > \theta_0$ based on a sample of n observations the expected evidence is $\sqrt{n}\, \mathcal{K}(\theta)$ for each θ. To attain a desired expected evidence τ_1 against alternative θ_1 one needs to choose n_1 to be the smallest integer greater than or equal to $\{\tau_1/\mathcal{K}(\theta_1)\}^2$.*

- K_2 Power calculations. *In the Neyman–Pearson setting the power function of a level α test based on T_n satisfies*

$$\begin{aligned}\Pi(\theta) &= P_\theta(T_n \geq z_{1-\alpha})\\ &= \Phi(\tau - z_{1-\alpha})\\ &= \Phi(\sqrt{n}\,\mathcal{K}(\theta) - z_{1-\alpha}).\end{aligned} \tag{17.4}$$

- K_3 Finding confidence intervals for θ. *A 100*$(1-\alpha)$ *% confidence interval for* θ *is given by*

$$\left[\mathcal{K}^{-1}\left(\frac{\{T_n - z_{1-\alpha/2}\}}{\sqrt{n}}\right),\ \mathcal{K}^{-1}\left(\frac{\{T_n + z_{1+\alpha/2}\}}{\sqrt{n}}\right)\right], \tag{17.5}$$

  where $\mathcal{K}^{-1}$ *is the inverse function to* $\mathcal{K}$.

It is tempting to interpret the Key Inferential Function as the expected evidence attainable with the statistical model and only $n = 1$ observation. However, usually $T_n \sim N(\sqrt{n}\,\mathcal{K}(\theta), 1)$ only for moderate to large sample sizes. It is our experience that the sample sizes required by these methods for inference are almost always smaller than those based on standard asymptotics using the central limit theorem alone.

Note that if the initial statistical model is reparametrized in terms of $\eta = m(\theta)$, where m is a strictly increasing function, then the Key Inferential Function remains unchanged; that is, after variance stabilization the resulting $T_n \sim N(\sqrt{n}\,\mathcal{K}(m^{-1}(\eta)), 1)$, where $\mathcal{K}$ is the function described above.

Another caveat is that the simplicity suggested by properties K_1 to K_3 is not always available; in particular if the underlying test statistics have a skewed distribution under both null and alternatives, it may not be possible to find a *vst* whose expectation satisfies $\tau \doteq \sqrt{n}\,\mathcal{K}(\theta)$ to a useful degree, for a $\mathcal{K}$ that is free of the sample size n.

A notable advantage of the methodology based on *vst*s is that it facilitates the comparison and/or combination of evidence from several related studies, because the evidence from all studies is placed on the same calibration scale.

17.3 Poisson model example

The theory of Section 17.2 is applied to counts data from one-sample experiments for which the Poisson model is appropriate. The test statistic is moved onto the canonical scale by a simple variance stabilizing transformation and interpreted as evidence, which leads to confidence intervals for the unknown model parameter. Later chapters will follow and extend the same methodology.

17.3.1 Example of counts data

A standard method for ascertaining the concentration μ of cells growing in a culture is to place a square grid over the region and count the number of cells in each of n randomly chosen squares of unit area. If the cells are randomly distributed over the region, it is reasonable to assume the resulting numbers $X_1, \ldots, X_n$ are independent, each with the same Poisson(μ) distribution:

$$P(X_i = x) = \mathrm{e}^{-\mu}\,\frac{\mu^x}{x!}\,, \text{for } x = 0, 1, 2, \ldots \tag{17.6}$$

It is well known that $\mathrm{E}[X] = \mu = \mathrm{Var}[X]$, so μ is the mean number of cells per unit area. The standard error in estimating μ by the sample mean $\bar{X}_n$ is

$$\mathrm{SE}[\bar{X}_n] = \sqrt{\mathrm{Var}[\bar{X}]} = \sqrt{\frac{\mu}{n}} \approx \sqrt{\frac{\bar{X}_n}{n}}.$$

It is desired to know whether $\mu \le \mu_0$ or $\mu > \mu_0$. The usual test statistic is $\bar{X}_n$, and for large n the p-value of the test which rejects for large values of $\bar{X}_n$ can be computed as follows. Having observed $\bar{X}_n = \bar{x}_n$

$$p = P(\bar{X}_n \ge \bar{x}_n) = P\left\{\sqrt{\frac{n}{\mu_0}}(\bar{X}_n - \mu_0) \ge \sqrt{\frac{n}{\mu_0}}(\bar{x}_n - \mu_0)\right\} \approx \Phi(-z_0), \tag{17.7}$$

where $z_0 = \sqrt{n}\,(\bar{x}_n - \mu_0)/\sqrt{\mu_0}$. This p-value is based on an asymptotic approximation, so will be referred to as p_{asym}.

For example let $\mu_0 = 1$, and take three samples of sizes $n_1 = 10, n_2 = 25, n_3 = 100$ from the Poisson(μ) model. Assume these samples have respective sample means $\bar{X}_{n_1} = 3$, $\bar{X}_{n_2} = 1.2$, $\bar{X}_{n_3} = 1.4$. Then the approximate p-values given by (17.7) are respectively 0.000000, 0.158655 and 0.000032. The exact p-values to six decimal places are obtained using the fact that $n\bar{X}_n$ has the Poisson$(n\mu_0)$ distribution and are 0.000001, 0.182140 and 0.000092. The extremely small p-values will be hard to interpret, whether one computes precise probabilities or not, because one has almost no experience with such rare events. We want to measure the evidence on the canonical scale, instead of trying to interpret the p-value, so we need to transform the test statistic.

17.3.2 A simple *vst* for the Poisson model

The standard test statistic is the sample mean $\bar{X}_n$ of n observations from the Poisson(μ) distribution. Now $\mathrm{Var}[\bar{X}] = g(\mathrm{E}[\bar{X}])$ for $g(t) = t/n$, so by (17.3) a possible choice for h is $h(x) = \sqrt{n}\int^x t^{-1/2}\,\mathrm{d}t = \sqrt{4nx}$. This heuristic argument suggests $Y = \sqrt{4n\bar{X}_n}$ will have variance approximately 1. However, this approximation cannot hold for all μ, because for fixed n, as μ approaches 0, so also will $\bar{X}_n$. Thus while $\mathrm{Var}[Y]$ may be

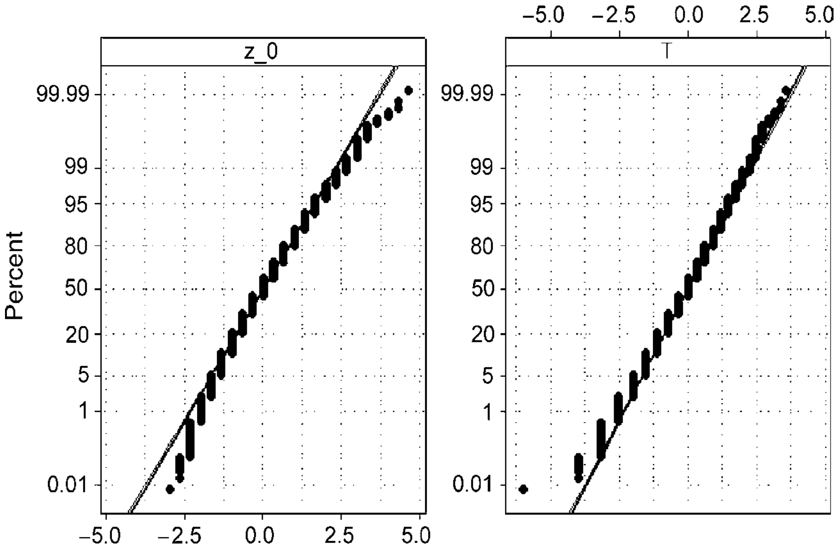


Figure 17.1 The left hand panel shows a probability plot of 10 000 values of $z_0 = 3(\bar{X}_9 - 1)$, where $\bar{X}_9$ is the mean of nine observations from the Poisson distribution with mean 1. The right hand panel shows a similar plot for the variance stabilized values $T = 6(\sqrt{\bar{X}_9} - 1)$. These plots show that the distribution of T is much closer to normality than z_0, especially in the upper tail, where the p-values are computed for testing $\mu = 1$ against $\mu > 1$. See Table 17.1.

approximately 1 for all μ not too near 0, we need to know when the approximation is good; and, as explained further below, a good 'rule of thumb' is $n\,\mu \geq 5$.

Next consider the expected value of the transformed test statistic Y. The first term in (17.1) gives the approximate mean $\mathrm{E}[Y] \doteq \sqrt{4n\mu}$. Further, the second derivative of h is $h''(x) = -\sqrt{n/4x^3}$, so the second term $-1/\sqrt{16n\mu}$ is the *bias term*. When $n\,\mu \geq 5$, the ratio of the bias term to the first term in (17.1) is $1/8n\mu \leq 1/40$.

The above results suggest that for testing $\mu = \mu_0$ versus $\mu > \mu_0$ the evidence be defined by $T = h(\bar{X}) - h(\mu_0)$, for then the resulting T would have approximate mean $\tau = \mathrm{E}[T] \doteq 2\sqrt{n}\,\{\sqrt{\mu} - \sqrt{\mu_0}\}$ and variance $\mathrm{Var}[T] \doteq 1$. Further, the distribution of T is approximately normal (see Figure 17.1) for even small sample sizes n, so it satisfies the properties E_1 to E_4 of a measure of evidence for testing $\theta = 0$ against $\theta > 0$, where $\theta = \mu - \mu_0$ is the effect. The evidence T can also be regarded as an estimator of its expectation τ, with $\mathrm{SD}[T] \doteq 1$. The *Key Inference Function* for this model is therefore $K(\mu) = 2\{\sqrt{\mu} - \sqrt{\mu_0}\}$.

It is instructive to examine the mean and variance of T as a function of μ, for the same example of $\bar{X}_9$, the mean of nine observations from a Poisson(μ) model. To illustrate the computation of evidence for the counts data of Section 17.3.1, we simply find $T_k = \sqrt{4n_k}\,\{\sqrt{\bar{X}_{n_k}} - 1\}$ for $k = 1, 2$ and 3. The results are shown in Table 17.2, along with the corresponding p-values $\Phi(-T_k)$. This tabulation shows that $\Phi(-T_k)$

Table 17.1 For selected values of μ the mean and standard deviation of the evidence $T = 6(\sqrt{\bar{X}_9} - 1)$ against $\mu = 1$ in favor of $\mu > 1$; the values shown are sample means and sample standard deviations based on 10 000 simulated samples of size 9 from the Poisson(μ) distribution. Note that the variance is quite stable, while the expected evidence τ rises with μ, as expected, with $\tau = 6(\sqrt{\mu} - 1)$.

μ	1	2	3	4	5	10	20
$\mathrm{E}_\mu[T]$	–0.08	2.42	4.34	5.96	7.37	12.95	20.82
$\mathrm{SD}_\mu[T]$	1.02	1.01	1.01	1.00	1.00	1.01	1.01

is a better approximation to the exact p-values than the asymptotic approximation p_{asym}. Both could be improved with continuity corrections. But the reader may well ask, what is the point of seeking approximations to very small p-values? For how can one possibly interpret them as measures of surprise, much less evidence, when one has no experience with evaluating such rare events?

On the other hand, there is no pretence to precision for the evidence values T_k shown in column 6 of Table 17.2; they each have standard error 1 in estimating their respective means, whether the null or alternative hypothesis is true. They give realistic assessments of the evidence against the null $\mu = 1$ in favor of the alternative $\mu > 1$. Because T_1 is almost 5, it is strong evidence against the null, while T_3 is moderate evidence and T_2 is almost negligible.

The reader may observe that the real issue is how to estimate the effect $\mu - 1$. This is indeed one of our goals. But if a 95 % confidence interval is found for μ in each of the three studies, they are roughly $\bar{X}_n \pm 2\sqrt{\bar{X}_n/n}$, or, respectively, [1.85, 4.15], [0.76, 1.64] and [1.16, 1.64]. The first interval appears to be estimating a different parameter than the second and third. Denoting the mean concentration in the kth sample by μ_k, in retrospect it would have been wiser to allow the means to be different. The strong assumption of homogeneity $\mu_1 = \mu_2 = \mu_3$ needs to be examined up front, and that is the purpose of Chapter 24. If this assumption is not tenable, then the analysis becomes more complicated when one wants to combine evidence in the three studies. One needs to decide what joint alternative in terms of the μ_k's is of interest, and choose an appropriate combination of the T_k's as evidence for it.

Table 17.2 Illustration of computations required for finding the evidence in three samples against $\mu = 1$ in favor of $\mu > 1$. The p-values at the right are exact to six places. The large-sample p-values $p_{\text{asym}} = \Phi(-z_0)$, where $z_0 = \sqrt{n_k}\,(\bar{X}_{n_k} - 1)$.

Sample k	n_k	$\bar{X}_{n_k}$	z_0	p_{asym}	T_k	$\Phi(-T_k)$	Exact p-value
1	9	3	6	0.000000	4.39	0.000006	0.000001
2	25	1.2	1	0.158655	0.95	0.169928	0.182140
3	100	1.4	4	0.000032	3.66	0.000124	0.000092

17.3.3 A better *vst* for the Poisson model

We showed the square root transformation could be applied to counts data which followed the Poisson model to obtain a stable variance. It was also seen that transformation led to approximate normality in our example. However, there are better transformations available if one wants variance stabilization, and these are not necessarily the same as those which transform towards normality. The problem of trying to obtain both desirable properties simultaneously for the Poisson model is a hard one which has been studied in some depth by Anscombe (1948), Efron (1982) and Bar-Lev and Enis (1988).

Anscombe (1948) found that $\sqrt{4n(\bar{X}_n + 3/8n)}$ would both reduce bias and stabilize the variance over a larger range of values of μ than $\sqrt{4n\bar{X}_n}$. Some results based on 100 000 simulations are shown in Figure 17.2.

17.3.4 Achieving a desired expected evidence

We conclude this section by choosing the sample size to achieve a desired expected evidence τ. For testing $\mu = \mu_0$ versus $\mu > \mu_0$ the evidence was defined by $T = 2\sqrt{n}\,\{\sqrt{\bar{X}_n} - \sqrt{\mu_0}\}$. The resulting $\tau = \mathrm{E}[T] \doteq 2\sqrt{n}\,\{\sqrt{\mu} - \sqrt{\mu_0}\}$ and variance $\mathrm{Var}[T] \doteq 1$. Therefore to obtain expected evidence τ for the alternative hypothesis when $\mu = c\mu_0$, where $c > 1$, take $n = \tau^2/4\mu_0(\sqrt{c} - 1)^2$. To be specific, let $\mu_0 = 1$ and $\mu = 4 = 4\mu_0$. Then the required $n = \tau^2/4$. If we want 'strong' expected evidence (which is three times the magnitude of 'weak' evidence 1.645) for this alternative when it is true, then we require $n = 6.25 \approx 6$.

17.3.5 Confidence intervals

We want a confidence interval for the effect $\theta = \mu - \mu_0$. It is easier to find the confidence interval for μ and then shift it to the left by μ_0. We can use the results already obtained, namely $\sqrt{4n\bar{X}_n} \sim N(\sqrt{4n\mu}, 1)$, for $n\mu$ sufficiently large, so in principle a $1-\alpha$ confidence interval for the mean $\sqrt{4n\mu}$ is centered on $\sqrt{4n\bar{X}_n}$, with length $2z_{1-\alpha/2}$. By dividing the endpoints of this interval by $\sqrt{4n}$ and then squaring them one obtains the confidence interval for μ:

$$\left[\left\{\sqrt{\bar{X}_n} - \frac{z_{1-\alpha/2}}{\sqrt{4n}}\right\}^2, \left\{\sqrt{\bar{X}_n} + \frac{z_{1-\alpha/2}}{\sqrt{4n}}\right\}^2\right]. \tag{17.8}$$

For example an approximate 95 % confidence interval for μ when $n = 9$, $\bar{X}_9 = 3$ and we take $z_{0.975} = 1.96 \approx 2$ is [1.96, 4.27], an interval which is not centered on 3 and is to be compared with the interval [1.85, 4.15] found earlier which is based on $\bar{X} \pm 2\,\mathrm{SE}[\bar{X}]$. We do not expect this last interval to have as accurate coverage as the one given by (17.8), because the standardized mean is not as close to normality as the variance stabilized mean (see Figure 17.1). An even better interval (in terms of accurate coverage probability) is obtained by using Anscombe's transformation. This amounts to replacing $\bar{X}_n$ by $\bar{X}_n + 3/8n$ in (17.8).

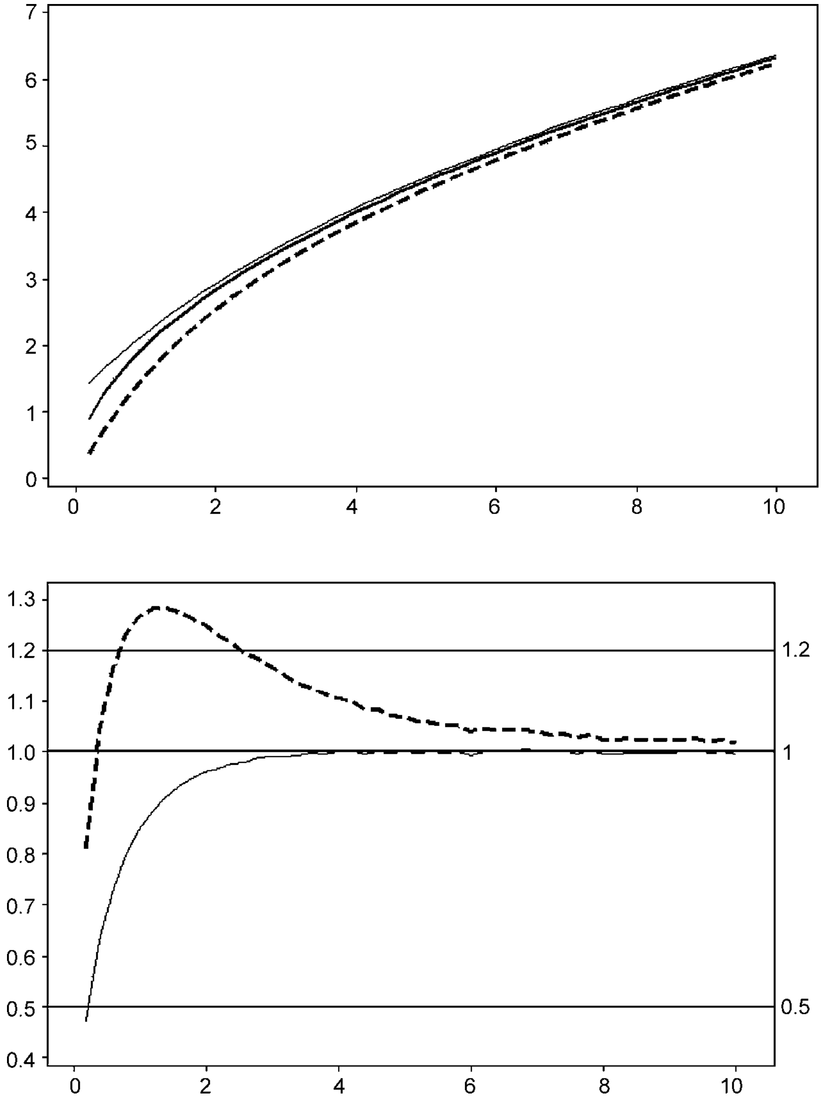


Figure 17.2 Let X ~Poisson(μ). The graph of $\text{SD}_\mu[\sqrt{4X}\,]$ versus μ is shown in the bottom plot as a dashed line, and $\text{SD}_\mu[\sqrt{4(X+3/8)}\,]$ as a thin solid line. In the top plot are shown the target $\sqrt{4\mu}$ versus μ as a thick solid line, and the expected values of the two transformed statistics $\text{E}_\mu[\sqrt{4X}\,]$ and the Anscombe *vst* $\text{E}_\mu[\sqrt{4(X+3/8)}\,]$ introduced in Section 17.3.3 versus μ as thick dashed and thin solid lines, respectively. For n observations, one can replace X by $n\bar{X}_n$ and μ by $n\mu$ on the horizontal axis. Note that for $n\mu \geq 5$, the Anscombe *vst* performs very well.

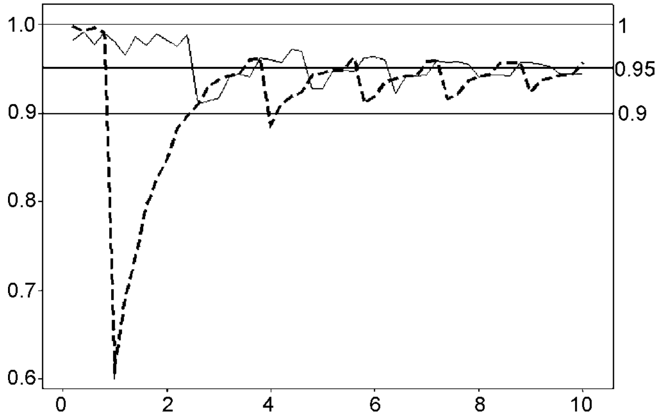


Figure 17.3 The empirical coverage probabilities of the nominal 95 % confidence intervals for μ based on (17.8), as a function of μ, for sample size 1, are shown as the dashed line. The thin solid line gives the same coverages for (17.8) with Anscombe's adjustment. See text for details.

17.3.6 Simulation study of coverage probabilities

While it is clear that the Anscombe *vst* stabilizes the variance and reduces the bias better than the simple square root *vst*, it is not clear that it leads to better 95 % confidence intervals. A simulation study based on 100 000 observations from the Poisson(μ) distribution were generated for μ ranging from 0.2 to 10 in steps of 0.2. For each observation the empirical coverage frequencies of (17.8) and the intervals based on Anscombe's *vst* were calculated; the results are shown in Figure 17.3. The coverage probabilities continue to stabilize around 95 % for larger μ, although this is not shown.

The main conclusion is that if one uses the Anscombe-based intervals, then the coverage probability always lies between 90 and 100 %, and for $\mu \geq 7$ the coverage is between 94 and 96 %. These results are for intervals based on 1 observation. For $n > 1$ observations it follows from the fact that $S_n = n\bar{X}_n$ is one observation with S_n ~Poisson(n, μ) that the coverages will be close to 95 % for $n\mu \geq 7$.

17.4 Two-sided evidence from one-sided evidence

We return to the basic model of Chapter 16.1, normal with unknown mean μ and known variance σ_0^2. For testing $\mu = \mu_0$ against the one-sided alternative $\mu > \mu_0$ the test statistic is the sample mean $\bar{X}_n$ and the measure of evidence for the one-sided alternative satisfying properties E_1 to E_4 of Section 16.1.1 is defined by $T = \sqrt{n}\,(\bar{X}_n - \mu_0)/\sigma_0$. Without loss of generality we can take $\mu_0 = 0$ and $\sigma_0 = 1$, so the measure of evidence reduces to $T = \sqrt{n}\,\bar{X}_n \sim N(\sqrt{n}\,\mu, 1)$. For simplicity of presentation, we only consider the case $n = 1$.

The problem now is to find a measure of evidence for the two-sided alternative $\mu \neq 0$, that is, $|\mu| > 0$. From the symmetry of the problem it is clear that the evidence

$T^{\pm}$ for the two-sided alternative should be be a function of the sufficient statistic $|T|$. Moreover, to be consistent with the definition of evidence for one-sided alternatives, $T^{\pm}$ should also satisfy properties E_1 to E_4 where the parameter of interest is now $|\mu|$.

We describe two very different methods leading to similar solutions. The first, in Section 17.4.1, is derived from the connection between $|\,T|$, which has a folded normal distribution, and the equivalent test statistic T^2, which has the noncentral chi-squared distribution with one degree of freedom. The second, in Section 17.4.2, is derived from doubling the *p*-value.

Both solutions yield positive values of $T^{\pm}$ for large $|\,T|$ that can be interpreted at evidence for the alternative $|\mu| > 0$. And both yield negative values of $T^{\pm}$ for small $|\,T|$ whose magnitude can be interpreted as positive evidence for the null $|\mu| = 0$. Thus the sign of $T^{\pm}$ is a signal as to which hypothesis is being supported: if negative, the null; if positive, the alternative. The magnitude $|\,T^{\pm}|$ gives the degree of evidence.

17.4.1 A *vst* based on the chi-squared statistic

The test of $|\mu| = 0$ against $|\mu| > 0$ that rejects the null for large $|T|$ is equivalent to the one that rejects for large $S = T^2$, and this statistic has the noncentral chi-squared distribution with one degree of freedom and noncentrality parameter $\lambda = \mu^2$. In symbols, $S \sim \chi^2_1(\lambda)$. For more information, see Chapter 22, wherein a *vst* for the general noncentral chi-squared distribution is derived and defined by (22.1). Now given a *vst* for any statistic, in this case S, a *vst* for a smooth one-to-one function of the statistic, in this case $|\,T| = +\sqrt{S}$, is the original *vst* composed with the inverse function, in this case the squaring function. Hence we are led to the following definition.

Definition 17.3 *Let F_1 be the cdf of the central chi-squared distribution with one degree of freedom and let c^2 be one-half the median of this distribution; i.e. $F_1(2c^2) = 0.5$. For the model $T \sim N(\mu, 1)$ and hypotheses $|\mu| = 0$ against $|\mu| > 0$, the two-sided evidence in $|\,T|$ is defined by*

$$T^{\pm} = \begin{cases} +\sqrt{S\ - c^2}\ - c, & \textit{for } S \geq 2c^2; \\ -\sqrt{S^* - c^2}\ + c, & \textit{for } S < 2c^2, \end{cases} \tag{17.9}$$

where $S^* = F_1^{-1}(1 - F_1(S))$. This definition assigns (negative) evidence to values of S less than the null median equal to the (positive) evidence assigned to corresponding values greater than the median, where the correspondence is in terms of tail area probabilities under the null; see also Definition 22.2.

The associated Key Inferential Function for the parameter $|\mu|$ is a special case of the Key for the noncentral chi-squared distribution given by Equation (22.3) for $N = 1$, $\nu = 1$ and $\theta = \lambda = \mu^2$. The median is $m_1 = 0.4549364$, so $c = 0.4769363$ and $\mathcal{L} = \mathcal{L}(\mu^2) = \sqrt{\mu^2 + 1 - c^2} = \sqrt{\mu^2 + 0.7725318}$ so

$$\mathcal{K}(|\mu|) = \mathcal{L} - \frac{(\mu^2 + 0.5)}{2\,\mathcal{L}^3} - c. \tag{17.10}$$

This $T^{\pm}$ satisfies property E_1 of Section 16.1.1 because it is a monotonically increasing function of the sufficient statistic $|\,T|$ (see the top plot in Figure 17.4).

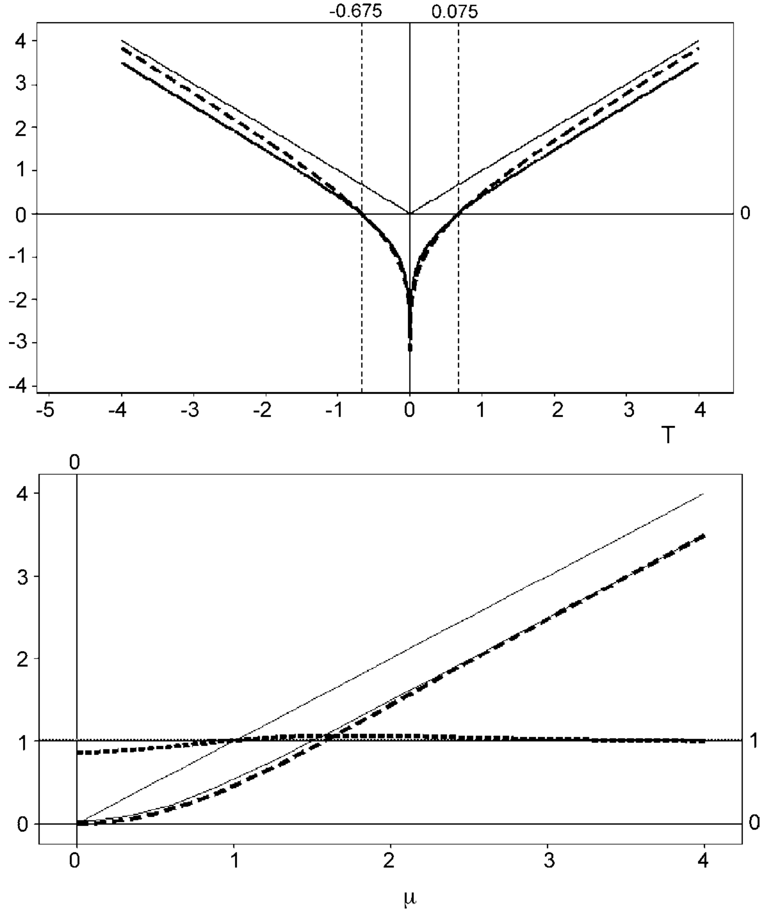


Figure 17.4 In the top plot is shown the graph of evidence for the two-sided evidence defined by (17.9) (the solid line) for comparison with the two-sided evidence suggested by doubling the one-sided p-value (17.11) (the dashed line). Also shown are reference lines depicting the graph of $|T|$ and $T = \pm 0.6745 = \Phi^{-1}(0.75)$. In the bottom plot is shown the empirical mean (dashed line) and standard deviation (dotted line) of $T^{\pm}$ defined in (17.9) based on 100 000 samples from the $N(\mu, 1)$ distribution for selected values of μ. The solid line is the graph of the Key Inferential Function defined by (17.10). By symmetry, the plot for negative μ is a reflection about the vertical axis. For interpretation, see the text.

The other properties were checked by experiments, in which the empirical means and standard deviations were based on 100 000 simulated values of $T^{\pm}$ at selected values of $|\mu|$. The bottom plot in Figure 17.4 suggests that $T^{\pm}$ satisfies property E_3, a stable standard deviation near 1. The estimated mean evidence E[$T^{\pm}$], also shown in the bottom plot of Figure 17.4 as a dashed line, grows from 0 with $|\mu|$, thus satisfying

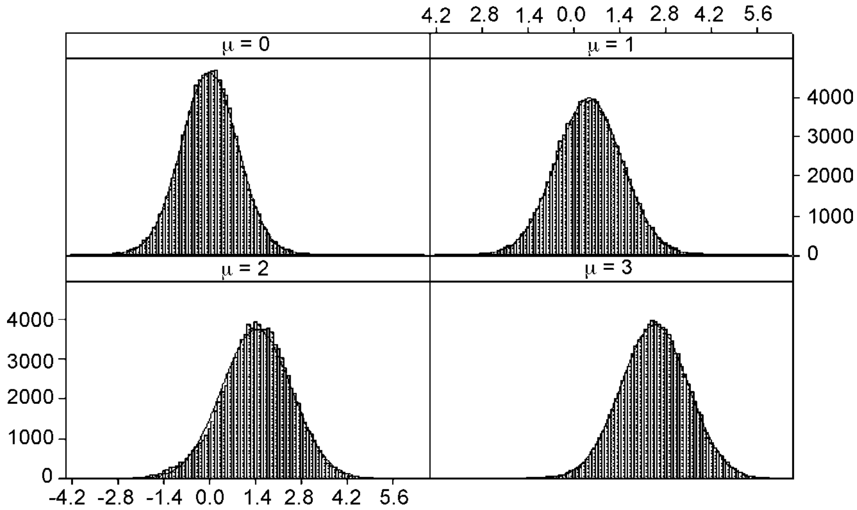


Figure 17.5 Histograms of $T^{\pm}$ defined by (17.9) based on 100 000 simulated values of $T \sim N(\mu, 1)$, for each of the four cases $\mu = 0, 1, 2$ and 3.

property E_4. Further, it closely approximates the Key Inferential Function (17.10), also plotted as a solid curve.

The histograms in Figure 17.5 summarize the results of 100 000 simulated values of $T^{\pm}$ for four cases $\mu = 0, 1, 2$ and 3, together with superimposed normal densities having the same mean and standard deviation. The first distribution is more concentrated because it has a standard deviation near 0.86, while the others have standard deviations closer to 1 (1.01, 1.06, 1.03, respectively). From these graphs and direct plots of the densities (not shown) it is clear that desirable property E_3, normality, will almost be satisfied.

17.4.2 A *vst* based on doubling the *p*-value

To motivate another definition of evidence for two-sided alternatives $\mu \neq 0$, we invoke the probit transformation of the p-value for the one-sided alternative; recall $T = \Phi^{-1}(1 - p)$ and $p = \Phi(-T)$. For the two-sided alternative $|\mu| > 0$, the p-value is $p^{\pm} = 2\Phi(-|T|)$. The relation $T = \Phi^{-1}(1 - p)$ suggests evidence in the two-sided p-value be defined by

$$T_{PV}^{\pm} = \Phi^{-1}(1 - 2\Phi(-|T|)) = \Phi^{-1}(F_1(S)) \ , \text{ where } S = T^2. \tag{17.11}$$

Thus one-sided evidence of 1.96 or -1.96 would become two-sided evidence 1.645. The graph of this function is shown as a dashed line in the top plot of Figure 17.4. It is very similar to that for $T^{\pm}$ defined earlier by (17.9). The advantage of $T_{PV}^{\pm}$ over $T^{\pm}$ is that it preserves the p-value interpretation: $\Phi(-T_{PV}^{\pm}) = p^{\pm}$, by definition. One advantage of $T^{\pm}$ over $T_{PV}^{\pm}$ is that it has variance stabilized closer to 1, and another is that it also provides a natural link to Chapter 22 on evidence in the chi-squared statistic.

For $|T| > \mu^* = \Phi^{-1}(0.75)$, the two-sided *p*-value is less than 0.5, and the evidence $T^{\pm}_{PV}$ inherits the 'penalty' property of *p*-values: if one assumes a two-sided alternative when there is prior knowledge to assume a one-sided alternative, one pays the penalty of a larger *p*-value than necessary. On the other hand, assuming a one-sided alternative when unjustified leads one to overstate the significane of the result. What is the relationship between the evidence as defined here under these two different assumptions regarding alternatives? The difference $| T| - T^{\pm}_{PV}$ is the amount of evidence one loses for assuming a two-sided alternative when there is enough prior knowledge to assume a known direction; it is also the amount by which the evidence is overstated, by assuming a one-sided alternative when there is no basis for it.

17.5 Summary

In this chapter we described some standard methods for variance stabilization, and showed how they can lead to a Key Inferential Function for testing a model parameter. This function supplements the *p*-value in much the same way that the power function supplements the significance level in Neyman–Pearson hypothesis testing.

We illustrated variance stabilization for the Poisson model and demonstrated how it could be used to find the evidence for the one-sided alternative $\mu > \mu_0$ to the null hypothesis $\mu = \mu_0$. The evidence in the sample mean $\bar{X}_n$ for the alternative $\mu > \mu_0$ was defined to be $T = 2\sqrt{n}\left\{\sqrt{\bar{X}_n} - \sqrt{\mu_0}\right\}$, or better, with $\bar{X}_n$ replaced by $\bar{X}_n + 3/8n$. This *vst* of $\bar{X}_n$ has variance approximately 1 for all $n\mu \geq 5$, and its expected value is $\tau \doteq 2\sqrt{n}\left\{\sqrt{\mu} - \sqrt{\mu_0}\right\}$. Knowing this allows us to interpret evidence as weak, moderate or strong, where 'weak' is essentially a 0.05 result under the null. It also enables us to choose a sample size to obtain a desired amount of experimental evidence. The variance stabilization of $\bar{X}_n$ is accompanied by increased normality as n increases for fixed μ. In particular, for $n\mu \geq 7$ the 95 % confidence intervals derived from Anscombe's *vst* are quite reliable, and not too bad for all μ.

Returning to the basic normal model with unknown mean and standard deviation 1, we also considered the problem of defining evidence $T^{\pm}$ for two-sided alternatives in terms of evidence for a one-sided alternative. We provided two solutions, the first based on the equivalence between the test based on $|T|$ and that based on T^2, which has a noncentral chi-squared distribution with one degree of freedom. The *vst* for $|T|$ is a simple function of that for T^2, which is a special case of the *vst* defined for an arbitrary noncentral chi-squared distribution found in Chapter 22. Thus the two-sided evidence function for normal tests links up with the chi-squared tests studied later. The second solution was based on *p*-value arguments, and led to a *vst* for which the two-sided evidence had similar performance characteristics to the first solution.

There is no all-purpose rule to tell us when we have achieved all the desirable properties E_1 to E_4 of a measure of evidence for one-sided alternatives; for each model this needs to be checked. On the other hand, a great deal of research has already gone into *vst* for standard models, so we will draw on this literature whenever possible. Efron (1982) provides a method for obtaining a transformation to normality.

18

One-sample binomial tests

In this chapter we find the evidence in one-sample binomial data using the methods introduced in the last chapter. They will be compared with some standard methods and shown to be competitive in terms of leading to reliable moderate to large sample confidence intervals. They will allow for easy interpretation of the evidence in test statistics. The inference is concerned with the parameter p, the probability of an event in a binomial setting which is often called risk. In turns out that larger sample sizes are required to obtain reliable results when the risk p is close to 0 or 1, and in particular we will require the sample size n to satisfy $np(1-p) \geq 5$ for values of p of interest. Many researchers in medical statistics prefer to think in terms of relative risk or odds ratio, rather than risk differences, so we reformulate the notion of evidence in these terms in Section 18.3; these concepts are further investigated in Chapter 19 in the two-sample setting.

18.1 Variance stabilizing the risk estimator

Let X have the $B(n, p)$ distribution, with $0 < p < 1$. When testing $p = p_0$ against $p > p_0$ we want to find the evidence for the alternative. We also want to find confidence intervals for p or the *effect* $p - p_0$. The usual test statistic $\hat{p} = X/n$ has mean $\mathrm{E}[\hat{p}] = p$ and variance $\mathrm{Var}[\hat{p}] = p(1-p)/n$, which varies with p, so we seek to transform $\hat{p}$.

The variance $\mathrm{Var}[X] = g(\mathrm{E}[X])$, where $g(t) = t(1-t)/n$, so by the methodology described in Section 17.2 a *vst* is $h(x) = \sqrt{n}\,\int^{x}\{t(1-t)\}^{-1/2}\mathrm{d}t = 2\sqrt{n}$ arcsin $(\sqrt{x}) + c$. The constant c is taken to be 0 in this classic transformation, and Anscombe (1948) has shown that $2\sqrt{n}$ arcsin$(\sqrt{\tilde{p}}\,)$, where $\tilde{p} = (X+3/8)/(n+3/4)$, comes closer

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

to a normal distribution with unit variance and mean $2\sqrt{n}\ \arcsin(\sqrt{p})$ than does the transformation applied to $\hat{p}$.

Let $c{=}-2\sqrt{n}\ \arcsin(\sqrt{p_0}\,)$. Then $T{=}h(\tilde{p}){=}2\sqrt{n}\ \left\{\arcsin\left(\sqrt{\tilde{p}}\right) - \arcsin(\sqrt{p_0}\,)\right\}$ will have approximate mean $\tau(p) = \mathrm{E}_p[T]$ increasing from 0 as p increases from p_0, and thus satisfy property E_4, Section 16.1.1, of a measure of evidence. It should also satisfy properties E_1 to E_3 for a useful range of parameters n and p because it is only a shift of the Anscombe (1948) statistic.

Definition 18.5 *The Key Inferential Function for the binomial model when testing $p = p_0$ against $p > p_0$ is for each $p \geq p_0$ given by*

$$\mathcal{K}(p) = 2\{\arcsin(\sqrt{p}\,) - \arcsin(\sqrt{p_0}\,)\}. \tag{18.1}$$

For testing in the other direction $p = p_0$ against $p < p_0$ the appropriate Key would be the negative of (18.1).

For an example we took $p_0 = 0.0$, so $\mathcal{K}(p) = 2\arcsin(\sqrt{p}\,)$, and then simulated 400 000 values of T at each of $n = 9, 15, 30$ and for p ranging from $p_0 = 0.01$ to 0.99 in intervals of 0.02. The empirical means of $T/\sqrt{n}$ are plotted against p in Figure 18.1, along with the target $\mathcal{K}(p)$. The bias is quite negligible over the range shown, especially compared to the standard deviation. For $n = 9$ the standard deviation of T is stable and near 1 for $0.2 < p < 0.8$, but descends to 0 as p approaches 0 or 1. The range of variance stability increases with the sample size.

These plots contain the information required for any choice of p_0. For example, if p_0 were 0.5 rather than 0.0 so the hypotheses were $p = 0.5$ against $p > 0.5$, then the plot for the standard deviations would be simply the portion of the lower plot to the right of 0.5. Similarly for the means, except that the values would be reduced by $2\arcsin(\sqrt{p_0}\,)$, in this case $2\arcsin(\sqrt{0.5}\,) = \pi/2$.

Extensive simulations (not shown) demonstrate that approximate normality of T holds provided $np(1-p) \geq 5$. Thus T cannot be considered a measure of evidence (satisfying criteria E_1 to E_4 of Section 16.1.1) unless this condition holds. In the next section we show the results of confidence intervals derived from T and compare them with a standard large-sample method.

18.2 Confidence intervals for p

A nominal 95 % confidence interval for the mean evidence $\tau(p)$ is $T \pm z_{0.975}$, so a nominal 95 % confidence interval for p is $\tau^{-1}(T \pm z_{0.975})$, or

$$\left[\left\{\sin\left(\arcsin\left(\sqrt{\tilde{p}}\right) - \frac{z_{0.975}}{2\sqrt{n}}\right)\right\}^2,\ \left\{\sin\left(\arcsin\left(\sqrt{\tilde{p}}\right) + \frac{z_{0.975}}{2\sqrt{n}}\right)\right\}^2\right]. \tag{18.2}$$

These intervals are far more reliable than $\hat{p} \pm z_{0.975}\sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p} = X/n$, as shown in Figure 18.2. In the top plot the 95 % confidence intervals based on variance stabilization have for $n = 9$ actual coverage which is too conservative for p

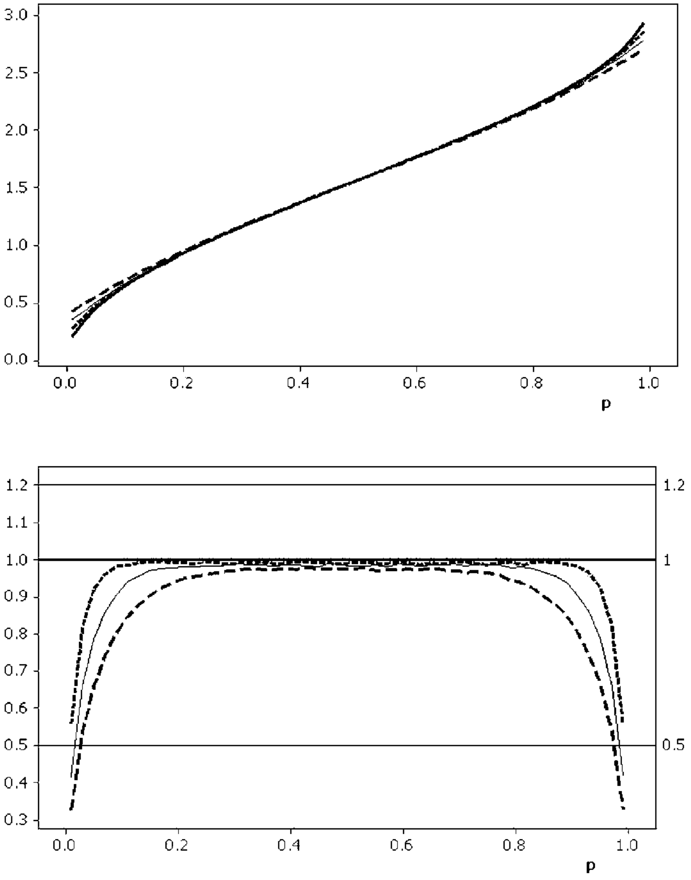


Figure 18.1 In this plot the null is taken to be $p = p_0 = 0$ and the alternative $p > 0$. Empirical means of $T/\sqrt{n} = 2\left\{\arcsin\left(\sqrt{\tilde{p}}\right)\right\}$ and standard deviations of T for $n = 9$, 15 and 30 are plotted as a function of p. The thick solid line in the upper plot is the graph of $\mathcal{K}(p) = 2\arcsin(\sqrt{p})$ versus p. The empirical means of $T/\sqrt{n}$ are shown for $n = 9$ as a dashed line, for $n = 15$ as a thin solid line and for $n = 30$ as a dotted line. In the lower plot are shown the corresponding empirical standard deviations for the same cases.

outside [0.25, 0.75], and too liberal for p near 0.4 and 0.6. The coverages for $n = 15$ are acceptable for p inside [0.25, 0.75], ranging from 93 to 97 % therein. For $n = 30$ the coverages are dependable for p inside [0.2, 0.8]. All suffer from spikes dropping below 95 % and overconservatism for p near 0 and 1.

In the bottom plot are shown the coverages of the classic large-sample intervals; all are much worse than those in the top plot. Even for $n = 30$ these intervals can descend to 88 % coverage. Replacing $\tilde{p}$ by $\hat{p}$ in any of these intervals only makes things worse.

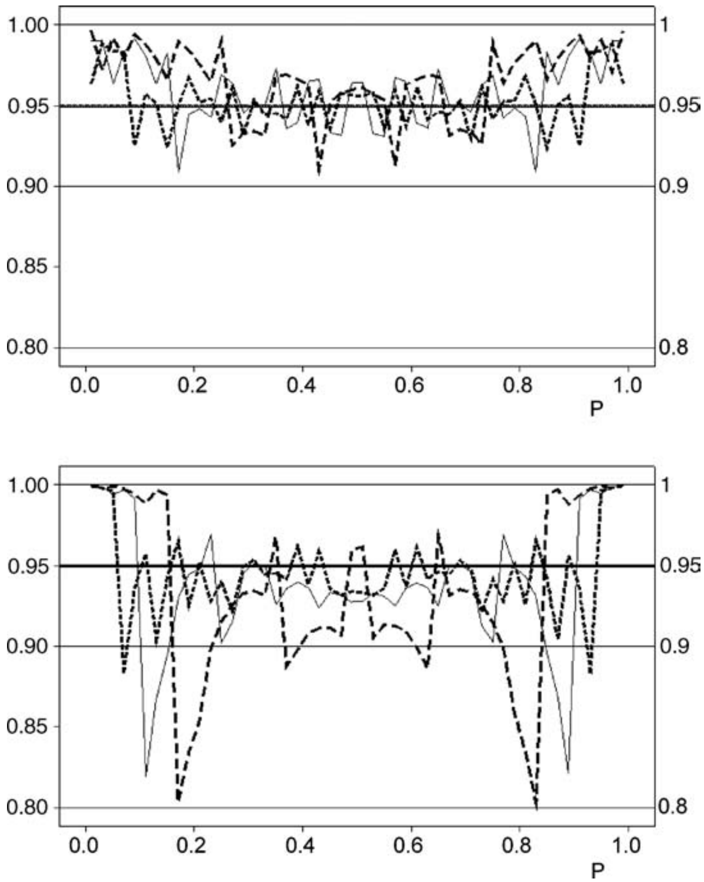


Figure 18.2 In the top plot are shown the empirical coverage probabilities of nominal 95 % confidence intervals based on (18.2) as a function of p for $n = 9$ as a dashed line, for $n = 15$ as a thin solid line and for $n = 30$ as a dotted line. By way of comparison the lower plot shows similar empirical coverage probabilities based on the large-sample confidence interval $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$.

For p near 0 or 1 larger sample sizes will be required to obtain evidence and/or reliable confidence intervals and this problem is examined in some detail in Section 18.4.

18.3 Relative risk and odds ratio

In comparing two risks p_1, p_2 many researchers prefer to think in terms of the *relative risk* $RR = p_1/p_2$ or *odds ratio* $OR = \{p_1/(1-p_1)\}\big/\{p_2/(1-p_2)\}$ rather than the risk difference $\Delta = p_1 - p_2$. When both risks are small the simple log transformation of estimators of the former two quantities has an approximate normal distribution, for sufficiently large samples sizes. In other words, a straightforward methodology is available for finding confidence intervals for these quantities (see pp. 24–28 of Lachin (2000), for example). We will compare these standard methods with other methods in Chapter 19. However, we can learn something about the log-transformed estimators of RR and OR by first considering the one-sample problem in which $X \sim B(n, p)$ and we are comparing an unknown risk p with a known null hypothesis value p_0. In this case only one parameter needs to be estimated.

18.3.1 One-sample relative risk

Consider the null hypothesis $p = p_0$ and alternative $0 < p < p_0$, which arises when p_0 is the 'known' risk in a certain population (risk of disease, positive response to a standard treatment, etc.), and p is the expected risk to a treated patient in the study. Rather than the simple difference $p_0 - p$ we are interested, say, in the relative risk $\mathrm{RR} = p_0/p$ which will exceed 1 under the alternative hypothesis, because then the treatment reduces the risk. Let $\theta = \ln(p_0/p)$. Inference for θ is equivalent to inference for $\mathrm{RR} = p_0/p$: the null hypothesis is now $\theta = 0$; the alternative $\theta > 0$.

Let X denote the number of positive responses in a study of n treated patients, and estimate p by $\hat{p} = X/n$. Then standard asymptotics shows that $\hat{\theta} = \ln(p_0) - \ln(\hat{p})$ has for increasing n an approximate normal distribution with mean θ and variance $(1 - p)/(np)$. It is customary to form a $100(1 - \alpha)$ % confidence interval $[L, U]$ for θ by taking $L = \hat{\theta} - z_{1-\alpha/2}\{(1 - \hat{p})/(n\hat{p})\}^{1/2}$ and similarly for U. Coverage can be slightly improved by modifying $\hat{p} = X/n$ to $(X + 0.5)/(n + 0.5)$ or $\tilde{p} = (X + 0.375)/(n + 0.75)$. A 95 % confidence interval $[L, U]$ for θ is easily transformed to a 95 % confidence interval $[e^L, e^U]$ for $\mathrm{RR} = p_0/p$.

The log-transformation is employed because when p is small the distribution of $\hat{p}$ is very skewed, while that of $\ln(\hat{p})$ is more symmetric. This raises the question of whether the evidence in the test statistic $p_0/\hat{p}$ for the alternative $p < p_0$ might be measured by the standardized transformed test statistic $\sqrt{n}\,\ln(p_0/\hat{p})[\{\hat{p}/(1 - \hat{p})\}]^{1/2}$. This statistic, for fixed $0 < p < 1$ and large enough n has an approximate normal distribution with variance 1 and asymptotic mean $\sqrt{n}\,\ln(p_0/p)\{p/(1 - p)\}^{1/2}$. This mean, when expressed in terms of the log-relative risk $\theta = \ln(p_0/p)$, is $\sqrt{n}\,K_0(\theta)$, with $K_0(\theta) = \theta\{p_0\mathrm{e}^{-\theta}/(1 - p_0\mathrm{e}^{-\theta})\}^{1/2}$. An example of this function for $p_0 = 0.2$ is plotted as a dashed line in Figure 18.3 for $\theta > 0$. Note that it cannot serve as a Key function because it is not monotonically increasing in θ over the range of alternatives. Also plotted is the Key function corresponding to the (negative of the) Key function in (18.1), namely $\mathcal{K}(p) = 2\sqrt{n}\,\{\arcsin(\sqrt{p_0}) - \arcsin(\sqrt{p})\}$, after reparametrization in terms of θ:

$$\mathcal{K}(\theta) = 2\big\{\arcsin(\sqrt{p_0}) - \arcsin(\sqrt{p_0\mathrm{e}^{-\theta}})\big\}. \tag{18.3}$$

The conclusion to be drawn from this comparison of $K_0(\theta)$ and $\mathcal{K}(\theta)$ is that while they both arise as *vst*s, the former the mean of the standardized $\hat{\theta}$, the latter the mean of the arcsine transformed $\hat{\theta}$, the former transformation has a limited range of applicability because it is not increasing in θ for all $\theta > 0$. And, as we will see, both

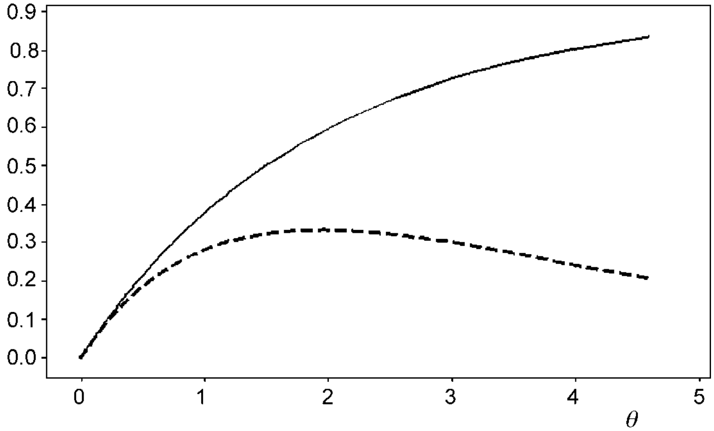


Figure 18.3 Plot of $K_0(\theta) = \theta\{p_0\mathrm{e}^{-\theta}/(1 - p_0\mathrm{e}^{-\theta})\}^{1/2}$ against $\theta = \ln(p_0/p)$ when $p_0 = 0.2$, with graph shown as a dashed line. Note that for small θ, corresponding to p near p_0, it is a close approximation to the Key function for the binomial model $\mathcal{K}(\theta) = 2\{\arcsin(\sqrt{p_0}) - \arcsin(\sqrt{p_0\mathrm{e}^{-\theta}})\}$, where again $p_0 = 0.2$, with graph shown as a solid line. Similar plots arise for different choices of p_0 over the unit interval, although both functions uniformly increase in θ with p_0, and the degree of approximation improves as p_0 approaches 0.

transformations achieve approximate normality at the same rate as n grows without bound. Therefore the arcsine transformation is our *vst* of choice.

18.3.2 One-sample odds ratio

Given the odds ratio $\mathrm{OR} = \{p_0/(1 - p_0)\}\big/\{p/(1 - p)\}$, relative to the null, let $\eta = \ln(\mathrm{OR})$. Then the original hypotheses $p = p_0$ versus $p < p_0$ can be reexpressed in terms of the odds ratio as $\mathrm{OR} = 1$ versus $\mathrm{OR} > 1$; or, if one prefers, $\eta = 0$ versus $\eta > 0$.

The estimator $\hat{\eta}$ obtained by substituting $\hat{p}$ for p in $\ln(\mathrm{OR})$ has for n increasing without bound a normal distribution with asymptotic mean η and variance $1/\{np(1 - p)\}$. That is, the standardized $\hat{\eta}$ has a limiting normal distribution with asymptotic mean $\{np(1 - p)\}^{1/2}\,\eta$ and variance 1. This raises the question as to whether the evidence in the log-odds ratio can be expressed in terms of the standardized $\hat{\eta}$ with Key function $\{p(1 - p)\}^{1/2}\ln[\{p_0(1 - p)\}/\{p(1 - p_0)\}]$. Equivalently, when reexpressed in terms of η, this possible mean evidence function becomes $K_1(\eta) = \eta\sqrt{c_0}\,\mathrm{e}^{\eta/2}/(1 + c_0\mathrm{e}^{\eta})$, where $c_0 = (1 - p_0)/p_0$. However, this function is not monotonically increasing over the alternatives $\eta > 0$, so it cannot serve as a Key function for a measure of evidence.

The negative of the Key function (18.1), which is appropriate for testing $p = p_0$ against $p < p_0$, can be reparametrized in terms of the log-odds ratio η to obtain a

Key function for $\eta = 0$ against $\eta > 0$:

$$\mathcal{K}(\eta) = 2\left\{\arcsin(\sqrt{p_0}\,) - \arcsin(\sqrt{1/(1+c_0\mathrm{e}^{\eta})}\,)\right\}. \tag{18.4}$$

This function is monotonically increasing in $\eta > 0$, and $K_1(\eta)$ defined above is a good approximation to it for small η. Again, as for the log-relative risk, the transformation to evidence is in terms of the classic arcsine transformation.

The main point is that if one prefers to think in terms of the simple difference $p_0 - p$ or the relative risk or the odds ratio the evidence in $\hat{p}$ remains the same. This evidence will be roughly normally distributed with variance 1 and mean function $\sqrt{n}$ times the same Key function, whether expressed in terms of the risk in (18.1) or the relative risk in (18.3) or the odds ratio in (18.4).

18.4 Confidence intervals for small risks *p*

It is clear from Figure 18.2 that the coverage probabilities of the confidence intervals for p examined earlier, one based on the standardized $\hat{p}$ and the other based on the *vst* arcsine transformation, vary greatly about the nominal 95 % value when p is near 0 or 1. The problem is caused by the highly skewed nature of the binomial distribution in these cases. Two additional methods are considered in this section, the log-transformation and a Poisson approximation to the binomial.

18.4.1 Comparing intervals based on the log and arcsine transformations

Again let $X \sim B(n, p)$, $\hat{p} = X/n$ and $\tilde{p} = (X + 0.375)/(n + 0.75)$. One possible remedy to the problem of asymmetry is to find a confidence interval $[L, U]$ for $\theta = \ln(p)$ using the normal approximation $N(\ln(p), (1 - p)/np)$ to the distribution of $\ln(\tilde{p})$. Then $[e^L,\ e^U]$ is taken to be the confidence interval for p. For nominal 95 % confidence, define $L = \ln(\tilde{p}) - z_{0.975}\{(1 - \tilde{p})/n\tilde{p}\}^{1/2}$, and $U = \ln(\tilde{p}) + z_{0.975}\{(1 - \tilde{p})/n\tilde{p}\}^{1/2}$. Another possibility is to use the arcsine-based confidence intervals defined earlier in (18.2). There is little to choose between the two methods for nominal 95 % confidence. As seen earlier in Figure 18.2, the empirical confidences zigzag about the nominal level. Extensive simulations with 40 000 simulations for sample sizes n ranging from 25 to 6400 and p ranging from 0 to 0.5 suggest that:

1. For $n = 25$ the log-transformed intervals had coverage between 93 and 97 % for $0.1 \le p \le 0.4$, while the arcsine-transformed intervals had the same range of coverage for $0.2 \le p \le 0.5$.

2. For $n \ge 50$ the log-transformed intervals described above have empirical confidence ranging from 93 to 97 % for p in the interval $[2.7/n,\ 0.5]$, while the arcsine intervals have the same range of empirical confidence over the intervals $[5/n,\ 0.5]$.

3. For $n \ge 100$ empirical coverage between 94 and 96 % is held over the smaller intervals $[11/n,\ 0.5]$ and $[19/n,\ 0.5]$, respectively, for the two methods.

To summarize, the confidence intervals derived by either method require increasingly large sample sizes as p approaches 0 in order for the empirical coverages to approach the nominal 95 % coverage. The log-transformation is more reliable over a slightly larger range of values of p for large n.

For p close to 1, one can use the $\ln(1 - \hat{p})$ transformation or the arcsine-based intervals in (18.2); by symmetry, the results will parallel those above, but with p replaced by $1 - p$.

A rule of thumb suggested by these simulations is that the when conditions $np(1 - p) \geq 5$ and $n \geq 25$ are satisfied, then the arcsine intervals will have empirical coverage between 93 and 97 %; and for $np(1 - p) \geq 11$ and $n \geq 100$, the coverages will lie between 94 and 96 %.

18.4.2 Confidence intervals for small *p* based on the Poisson approximation to the binomial

Another method uses the Poisson(μ) approximation to the $B(n, p)$ distribution with $\mu = np$. It is based on the result that as $n \to \infty$ and $p \to 0$ with $\mu = np$ held fixed, the random variable X converges in distribution to a variable $Y \sim$Poisson(μ). And for a Poisson variable we know that the classic square root transformation is an effective *vst*. In view of Figure 17.3 one can expect confidence intervals based on (17.8) applied to the single observation Y to have good coverage for all $\mu > 7$. These intervals for μ are of the form $\{Y \mp z_{0.975}/2\}^2$. Both endpoints need to be divided by n to obtain the interval for p. Empirical coverage probabilities of these intervals are, however, somewhat disappointing. This is explained as follows.

Decker and Fitzgibbon (1991) are cited in Johnson *et al.* (1993), p. 118; they recommend practical use of the Poisson approximation to the binomial when $n^{0.31}p < 0.47$. This bound, combined with our ‘rule of thumb’ that $\mu = np > 7$ to obtain good coverage of the Poisson parameter, suggests that these intervals will have reliable coverage when $7/n < p < 0.47/n^{0.31}$.

Even for sample size $n = 50$ this range of p is void, while for $n = 100$ it is $0.07 < p < 0.11$. For $n = 200$ it is $0.035 < p < 0.091$ and for $n = 500$ it is $0.014 < p < 0.068$. Extensive simulations for sample sizes ranging from 200 to 1000 reveal that the main advantage of these intervals over those based on the log-transformation, say, is that the coverage is more conservative and remains above 95 % for these intervals; the disadvantage is that this is so over only the narrow range of p just specified. Outside this range, but still within the interval $0.02 < p < 0.2$, they tend to be too conservative with coverages rising to 97 % or even 98 % even for $n = 1000$ and $p = 0.2$.

In summary, for small p we do not recommend the Poisson approximation to the binomial followed by the classic *vst* to obtain approximate large-sample confidence intervals unless the interest lies in values of p within the above narrow range of values, and at the same time outside a larger range of values of p where the log or arcsine transformations studied earlier yield reliable intervals for p for any given n.

18.5 Summary

In this chapter we found that the classic arcsine transformation of the test statistic $\hat{p}$ leads to statistical evidence T for testing $p = p_0$ against $p > p_0$. The expected evidence $\tau = \sqrt{n}\,\mathcal{K}(p)$ is determined by a Key Inferential Function $\mathcal{K}(p)$ that can be reparametrized in terms of the relative risk p_0/p or odds ratio $p_0(1 - p)/\{p(1 - p_0\}$. Minimum sample sizes were determined so that the transformation to evidence also led to reliable 95 % confidence intervals for p, even when it is near 0, and these intervals are easily converted to intervals for the relative risk or odds ratio.

19

Two-sample binomial tests

Let $X \sim B(n_1, p_1)$ independent of $Y \sim B(n_2, p_2)$. One parameter of interest is the (raw) *effect* $\Delta = p_1 - p_2$. When p_1, p_2 represent risks for control and treatment subjects, Δ is called the *risk difference*. In Section 19.1.1 we find a measure of evidence T for the alternative $\Delta > 0$ to the null hypothesis $\Delta = 0$. It turns out that the expected evidence is a simple function of a correlation effect size, which in turn is a monotonic function of the standardized effect. Minimal sample sizes required to obtain desired expected evidence for raw effects and effect sizes are found in Section 19.1.3. Then confidence intervals for these effect sizes are derived in Section 19.2, and in Section 19.3 are presented confidence intervals for the risk difference Δ. Other standard parameters of interest, especially when p_1, p_2 are small, are the *relative risk* $RR = p_1/p_2$ and the *odds ratio* $OR = p_1(1 - p_2)/\{p_2(1 - p_1)$. New and traditional methods for these parameters are discussed in Sections 19.1.4 and 19.4.

19.1 Evidence for a positive effect

19.1.1 Variance stabilizing the risk difference

Let $q = n_2/N$ represent the proportion of the total sample size $N = n_1 + n_2$ allotted to the second sample. Brown and Li (2005) introduce the parameter $p = qp_1 + (1 - q)p_2$ so the (p_1, p_2) unit square can be reparametrized in terms of Δ and p. Note that $p_1 = p + (1 - q)\Delta$ and $p_2 = p - q\Delta$.

The maximum likelihood estimators of p_1, p_2, Δ and p are $\hat{p}_1 = X/n_1$, $\hat{p}_2 = Y/n_2$, $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$ and $\hat{p} = q\hat{p}_1 + (1 - q)\hat{p}_2$, respectively. Brown and Li (2005) observe that $\text{Var}[\hat{\Delta}]$, when expressed in terms of p, q, Δ and N, is $\text{Var}[\hat{\Delta}] = (\zeta - \Delta^2)/N$, where $\zeta = \{p(1 - p)\}/\{q(1 - q)\}$ depends on the unknown parameter p.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte
© 2008 John Wiley & Sons, Ltd

Now for moderate and large sample sizes the distribution of $\hat{p}$ is approximately normal with mean p and variance

$$N\text{Var}[\hat{p}] = \{q^3 + (1-q)^3\}\zeta - (1-2q)(1-2p)\Delta - q(1-q)\Delta^2.$$

For equal sample sizes, $q = 0.5$, $p = \bar{p} = (p_1 + p_2)/2$ and the first term in this expression is $\bar{p}(1-\bar{p})$. The second term drops out, and the third term is less than $\bar{p}^2$ in magnitude. So for equal sample sizes the large-sample distribution of $\hat{p}$ does not depend (much) on Δ; that is, $\hat{p}$ is almost ancillary for Δ: it reveals little about Δ. We also note in passing that for $q = 0.5$, $\text{Var}[\hat{p}] = \text{Var}[\hat{\Delta}]/4$.

In view of these facts, it seems reasonable to stabilize the variance of $\hat{\Delta}$, conditional on $\hat{p} = p$. So we treat p, and hence ζ, as if they were known in the following paragraph, and then estimate them. We can expect the results to be useful for q near 0.5 and small Δ.

One can write $\text{Var}[\hat{\Delta}] = g(\text{E}[\hat{\Delta}])$, where $g(t) = a - bt^2$ with $a = \zeta/N$ and $b = 1/N$. Using the standard method described in Section 17.2, one obtains an indefinite integral $h(x) = \int^x |g(t)|^{-1/2}\text{d}t = b^{-1/2}\arcsin(x\zeta^{-1/2})$. This yields the conditional evidence, given $\hat{p} = p$, of $\sqrt{N}\arcsin(\hat{\Delta}\zeta^{-1/2})$. Then by substituting an estimate for p in ζ, a candidate for unconditional evidence is obtained.

Definition 19.1 *We found through experimentation that the choices* $\tilde{p}_1 = (X + 0.5)/(n_1 + 1)$, $\tilde{p}_2 = (Y + 0.5)/(n_2 + 1)$, *when substituted into the formulas for* Δ, p *and* $\zeta = \{p(1-p)\}/\{q(1-q)\}$, *lead to a measure of evidence*

$$T = \sqrt{N}\arcsin\left(\tilde{\Delta}\,\tilde{\zeta}^{-1/2}\right). \tag{19.1}$$

This T satisfies conditions E_1 to E_4 of a measure of evidence as defined in Section 16.1.1 for a wide range of parameter values Δ, p. It is clear that condition E_1 is satisfied because T is monotonically increasing in $\hat{\Delta}$. Approximate normality with unit variance will need to be checked by simulations, but condition E_4 is simpler to verify.

To this end, define the *standardized effect* $\delta = \Delta/\sqrt{N\,\text{Var}[\hat{\Delta}]} = \Delta/\sqrt{\zeta - \Delta^2}$. The associated *correlation effect size* is $\rho = \delta/\sqrt{1+\delta^2} = \Delta/\sqrt{\zeta}$, as shown in Section 1.3. Hence the first term in (19.1) for the expected evidence $\tau = \text{E}[T] \doteq \sqrt{N}\,\mathcal{K}(\rho)$ where $\mathcal{K}(\rho) = \arcsin(\rho)$. Note that this is monotonically increasing from 0 as Δ increases from 0; thus $\tau = \text{E}[T] \doteq \sqrt{N}\,\mathcal{K}(\rho)$ satisfies condition E_4 of a measure of evidence for testing $\Delta = 0$ against $\Delta > 0$.

Definition 19.2 *The Key Inferential Function for testing* $\Delta = 0$ *versus* $\Delta > 0$ *in the two-sample binomial model can be expressed in terms of the effect size as measured by* ρ *or* δ *and is given for each real* δ *by*

$$\mathcal{K}(\rho) = \arcsin(\rho)\,,\ \textit{where}\ \rho = \delta/\sqrt{1+\delta^2}. \tag{19.2}$$

19.1.2 Simulation studies

As an example, we examine the behavior of T for equal sample sizes $n_1 = n_2$ and fixed $p = 0.5$. In this case $\Delta = \rho$. The Key Inferential Function is shown as a thick solid line in the upper plot of Figure 19.1. Also plotted are three graphs of the empirical means of $T/\sqrt{N}$ versus Δ, where $N = n_1 + n_2 = 2n_1 = 18$, 30 and 60. All means are based on 100 000 simulations at Δ ranging from 0.01 to 0.99 in steps of 0.02. It is clear that the bias in $T/\sqrt{N}$ for estimating $\mathcal{K}(\rho)$ is decreasing for all Δ as the sample size increases and further it is negligible for $\Delta < 0.5$ and these sample sizes. The corresponding standard deviations of T are plotted as functions of Δ in the lower plot and one can see that variance stabilization is achieved for a wide range of Δ.

For a second example, we repeated the above experiments, but now with $q = 2/3$; that is, $n_2 = 2n_1$ and $N = 3n_1 = 18, 30$, and 60. The only difference in the results is that the alternative hypothesis is restricted to [0, 0.75] because in general

$$\max\{-p/(1-q), (p-1)/q\} \leq \Delta \leq \min\{p/q, (1-p)/(1-q)\}, \tag{19.3}$$

due to the restriction that (p_1, p_2) lies in the unit square. There is a slight lowering of the Key Inferential Function because $q \neq 0.5$, and the behavior of T defined by (19.1) is similar to that depicted in the plots in Figure 19.1. The main disadvantage of using unbalanced sampling is that actual and expected evidence is lower than it would be with balanced sampling. But for these sample sizes the variance stabilization works well when $p = 0.5$.

For the third and fourth examples we fixed $p = 0.2$ with the same total sample sizes as above and considered both cases $q = 0.5$ and $q = 2/3$. However, except for the total sample size of $N = 60$, the results are disappointing, and they suggest that for small p larger sample sizes are required for T defined by (19.1) to be useful as a measure of evidence.

19.1.3 Choosing sample sizes to achieve desired expected evidence

In order to attain expected evidence $\tau_1 \doteq \sqrt{N}\,\arcsin(\rho_1)$ for a correlation effect size ρ_1 one requires $N \geq \{\tau_1/\arcsin(\rho_1)\}^2$. In particular, to attain 'moderate' expected evidence of $3.3 = 2 \times 1.645$ for $\rho_1 = 0.5$, one needs a total sample size of $N \geq (6 \times 3.3/\pi)^2 = 39.7$, or $N = 40$. This could be apportioned equally, or into somewhat unequal samples whose sum is 40. Further below we consider some cases of unequal sample sizes.

To achieve an expected evidence of $\tau_1 \doteq \sqrt{N}\,\arcsin(\Delta_1/\sqrt{\zeta})$ against an effect Δ_1, it suffices, for any fixed q, to take $N \geq \{\tau_1/\arcsin(2\sqrt{q(1-q)}\,\Delta_1)\}^2$. In particular, for $\tau_1 = 3.3$ and $\Delta_1 = 0.5$, it suffices to take equal sample sizes totaling $N = 40$. For unequal sample sizes, a larger total is required.

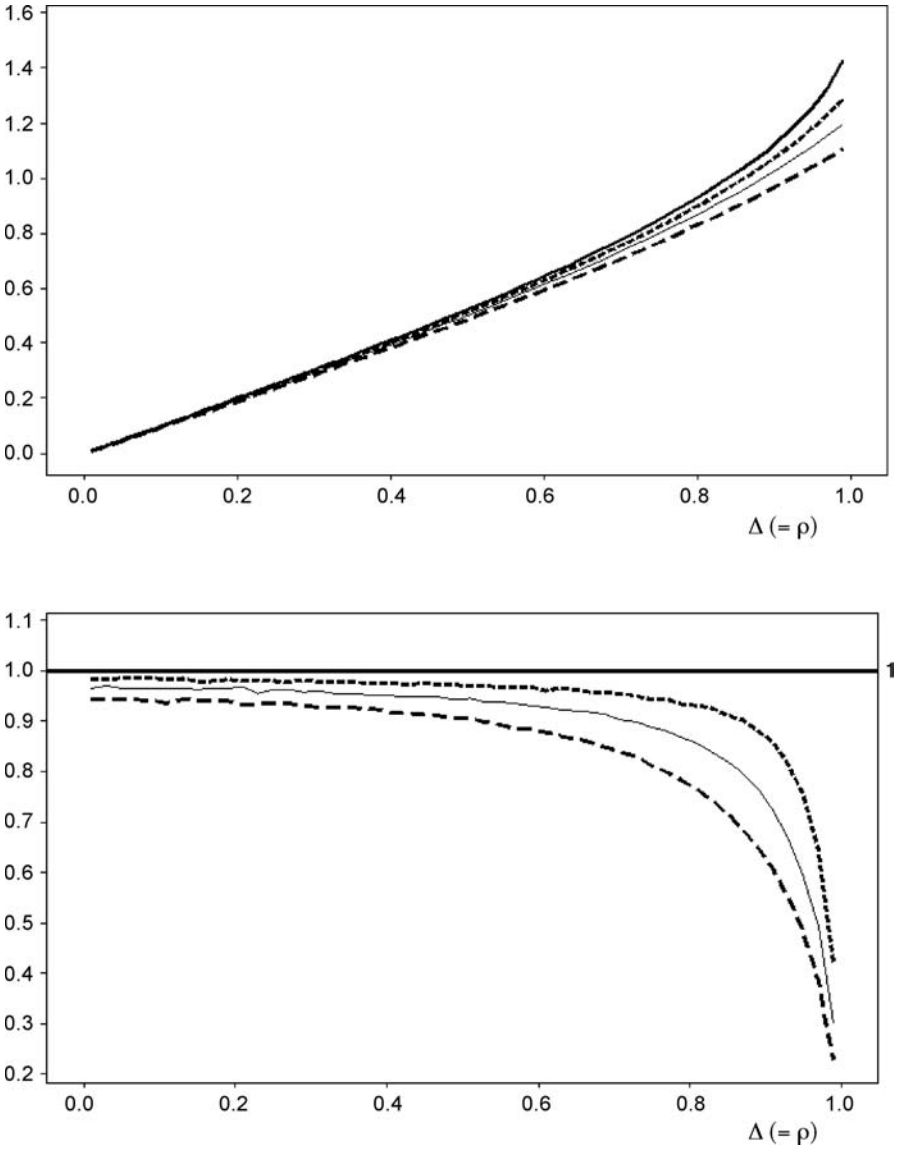


Figure 19.1 Empirical means of $T/\sqrt{n}$ and standard deviations of T for equal sample sizes plotted as a function of $\Delta = p_1 - p_2$. The parameter $p = (p_1 + p_2)/2$ is fixed at $p = 0.5$ so $\zeta = 1$ and $\Delta = \rho$. The thick solid line in the upper plot is the graph of $\mathcal{K}(\rho) = \arcsin(\rho)$ versus ρ. The empirical means of $T/\sqrt{N}$ are shown for $m = n = 9$ as a dashed line, for $m = n = 15$ as a thin solid line and for $m = n = 30$ as a dotted line. In the lower plot are shown the corresponding empirical standard deviations of T for the same cases.

19.1.4 Implications for the relative risk and odds ratio

Using the identities $p_1 = p + (1-q)\Delta$ and $p_2 = p - q\Delta$ one can rewrite the relative risk $\mathrm{RR} = (p + (1-q)\Delta)/(p - q\Delta)$ and see that for fixed p, it is strictly increasing in Δ because $0 < q < 1$. Similarly, the odds $p_1/(1-p_1) = (p + (1-q)\Delta)/(1 - p - (1-q)\Delta)$ is for fixed p strictly increasing in Δ, while the odds $p_2/(1-p_2)$ is strictly decreasing in Δ. Thus for fixed p the odds ratio $\mathrm{OR} = p_1(1-p_2)/\{(1-p_1)p_2\}$ is also strictly increasing in Δ.

The evidence $T = \sqrt{N}\ \arcsin\bigl(\tilde{\Delta}\tilde{\zeta}^{-1/2}\bigr)$ defined in (19.1) was derived by a conditional argument, given $\tilde{p} = p$, so this conditional evidence for $\Delta > 0$ can serve as evidence for $\mathrm{RR} > 0$ or for $\mathrm{OR} > 0$. Thus the evidence for a positive effect, whether it be parametrized by Δ, RR or OR, is the same. The simulation studies in Section 19.1.2 indicate that this evidence has good unconditional properties as well.

19.2 Confidence intervals for effect sizes

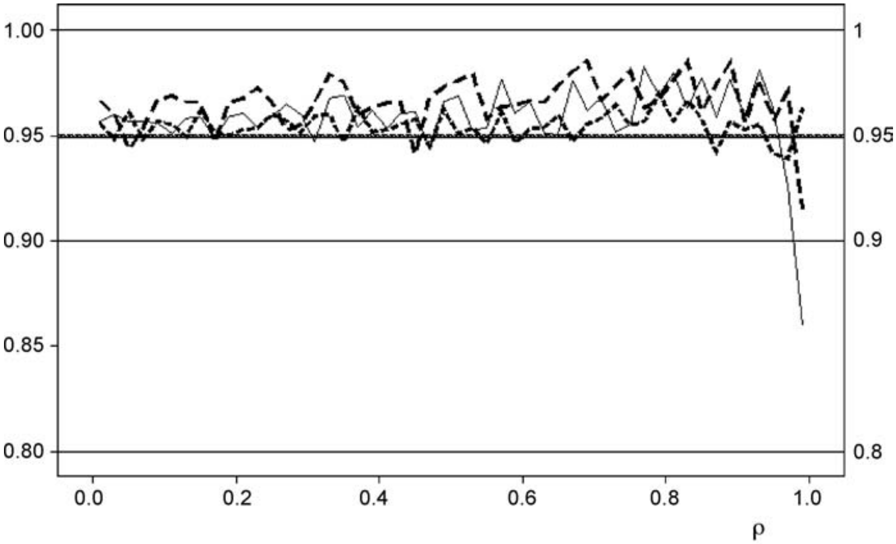
While it is clear that for large enough sample sizes n_1, n_2 the distribution of T defined by (19.1) will be approximately normal with asymptotic mean $\tau \doteq \sqrt{N}\,\mathcal{K}(\rho)$ and variance 1, so that $T \pm z_{0.975}$ will provide nominal 95 % confidence intervals for τ, simulation studies are required to determine how well these intervals perform. Any interval for τ is easily transformed into an interval $[L, U]$ for the correlation effect size ρ, with

$$[L, U] = \left[\sin\left(\frac{T - z_{0.975}}{\sqrt{N}}\right),\ \sin\left(\frac{T + z_{0.975}}{\sqrt{N}}\right)\right]. \tag{19.4}$$

And this leads immediately to intervals having the same confidence for the standardized effect δ, namely $[L/\sqrt{1-L^2}\ ,\ \ U/\sqrt{1-U^2}\ ]$. Note that the above preservation of intervals under transformations tacitly assumed that the argument of the sine function in the definition of L, U lies within the interval $[-\pi/2,\ \pi/2]$. This is the case for $|T| < 3.0787\sqrt{N}$.

In the top plot of Figure 19.2 are shown the empirical coverage probabilities of nominal 95 % confidence intervals $[L, U]$ for ρ as defined above based on 100 000 simulations. In the top plot $p = 0.5$ and there are three cases of equal sample sizes $n_1 = n_2$. For a total sample size of $N = 18$, the dashed line shows the empirical coverage ranges from 94.5 to 98 % for all ρ not too near 1. For a total $N = 30$, the thin solid line shows the coverages range from 95 to 97.5 % for the same ρ. For $\rho < 0.5$ the results tend to be closer to 95 % , but there is always some slight fluctuation in coverage. For $N = 60$ the empirical coverages continue to improve.

In the bottom plot of Figure 19.2 are shown similar results for $p = 0.2$ and equal sample sizes. Because p_1, p_2 are small, with average 0.2, larger sample sizes are required to get accurate converage for ρ. Even though we took $N = 30, 60$ and 120, only the last really leads to accurate coverage of 95 % confidence intervals, and then only for $\rho < 0.2$.



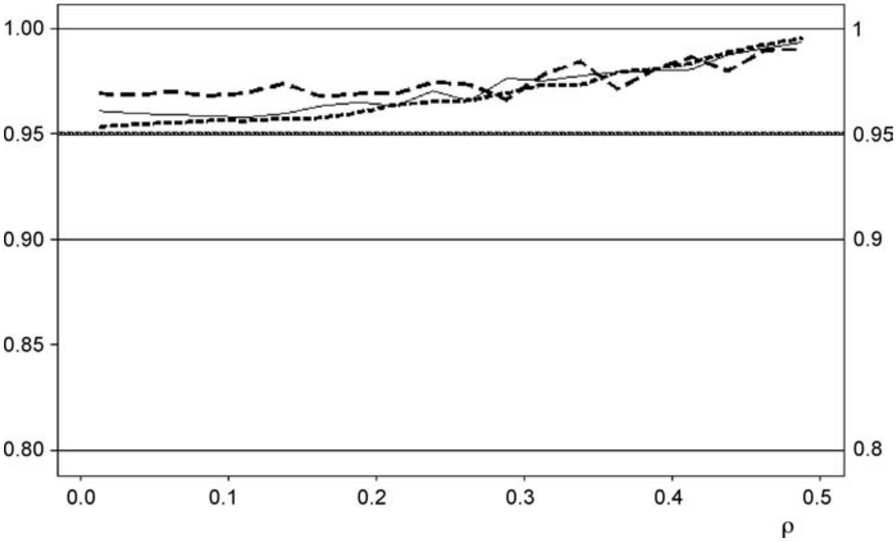


Figure 19.2 The upper plot shows empirical coverage probabilities of nominal 95 % confidence intervals for ρ defined by (19.4) when $p = 0.5$ and equal sample sizes totaling $N = n_1 + n_2 = 18$ (dashed line), $N = 30$ (thin solid line) and $N = 60$ (dotted line). The lower plot gives similar results when $p = 0.2$ and there are equal sample sizes totaling $N = 30$ (dashed line), $N = 60$ (thin solid line) and $N = 120$ (dotted line).

19.3 Estimating the risk difference

The nominal 95 % confidence intervals for $\rho = \Delta/\sqrt{\zeta}$ given by (19.4) can be multiplied by $\sqrt{\tilde{\zeta}}$, where $\tilde{\zeta}$ is an estimate of ζ, to yield nominal 95 % confidence intervals for Δ. They can be expressed as

$$\left[\sqrt{\tilde{\zeta}}\left(\frac{T - z_{0.975}}{\sqrt{N}}\right), \sqrt{\tilde{\zeta}}\left(\frac{T + z_{0.975}}{\sqrt{N}}\right)\right]. \tag{19.5}$$

Despite the additional estimate required, they tend to have better coverage properties than the corresponding intervals for ρ under the same conditions (see Figures 19.2. and 19.3).

19.4 Relative risk and odds ratio

We continue with the model $X \sim B(n_1, p_1)$ independent of $Y \sim B(n_2, p_2)$ where p_1, p_2 represent risks for control and treated subjects, respectively, and it is now desired to find a confidence interval for the relative risk or odds ratio. The methodology for finding confidence intervals in Sections 19.4.1 and 19.4.2 is standard and can also be found in Sections 2.3.2 and 2.3.3 of Lachin (2000), for example. New methods for finding confidence intervals for the relative risk and odds ratio are presented in Section 19.4.3.

19.4.1 Two-sample relative risk

Let the null hypothesis be no difference between treatment and control, with alternative that the treatment reduces the risk more than the control. The relative risk RR $= p_1/p_2$ which will exceed 1 under the alternative hypothesis. Let $\theta = \ln(p_1/p_2)$. Inference for θ is equivalent to inference for the RR : the null hypothesis is now $\theta = 0$; the alternative $\theta > 0$.

Let $\hat{p}_1 = X/n_1$, $\hat{p}_2 = Y/n_2$, $N = n_1 + n_2$ and $q = n_2/N$. Standard asymptotics shows that $\hat{\theta} = \ln(\hat{p}_1/\hat{p}_2)$ has, for large n_1, n_2, an approximate normal distribution with asymptotic mean θ and variance

$$\mathrm{Var}[\hat{\theta}] = \frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}. \tag{19.6}$$

This formula assumes $p_1, p_2 > 0$ and so we modify $\hat{p}_1 = X_1/n_1$ to $\tilde{p}_1 = (X_1 + 0.5)/(n_1 + 0.5)$ and similarly for $\hat{p}_2$. Using (19.6) the standard error of $\tilde{\theta} = \tilde{p}_1/\tilde{p}_2$ is estimated by $\mathrm{SE}[\tilde{\theta}] = \{(1-\tilde{p}_1)/(n_1\tilde{p}_1) + (1-\tilde{p}_2)/(n_2\tilde{p}_2)\}^{1/2}$.

One can then form a $100(1-\alpha)$ % confidence interval $[L, U]$ for θ by taking $L = \hat{\theta} - z_{1-\alpha/2}\,\mathrm{SE}[\hat{\theta}]$ and $U = \hat{\theta} + z_{1-\alpha/2}\,\mathrm{SE}[\hat{\theta}]$. This interval $[L, U]$ for θ is then transformed to a $100(1-\alpha)$ % confidence interval $[e^L, e^U]$ for RR.

The standardized estimator $S = (\hat{\theta} - 0)/\mathrm{SE}[\hat{\theta}]$; then as n_1, n_2 increase without bound, we have S converging in distribution to an $N(\mathrm{E}[S], 1)$ distribution, with asymptotic mean $\mathrm{E}[S] \doteq \sqrt{q(1-q)N}\;\theta/\{qp_2(1-p_1) + (1-q)p_1(1-p_2)\}^{1/2}$. It is not

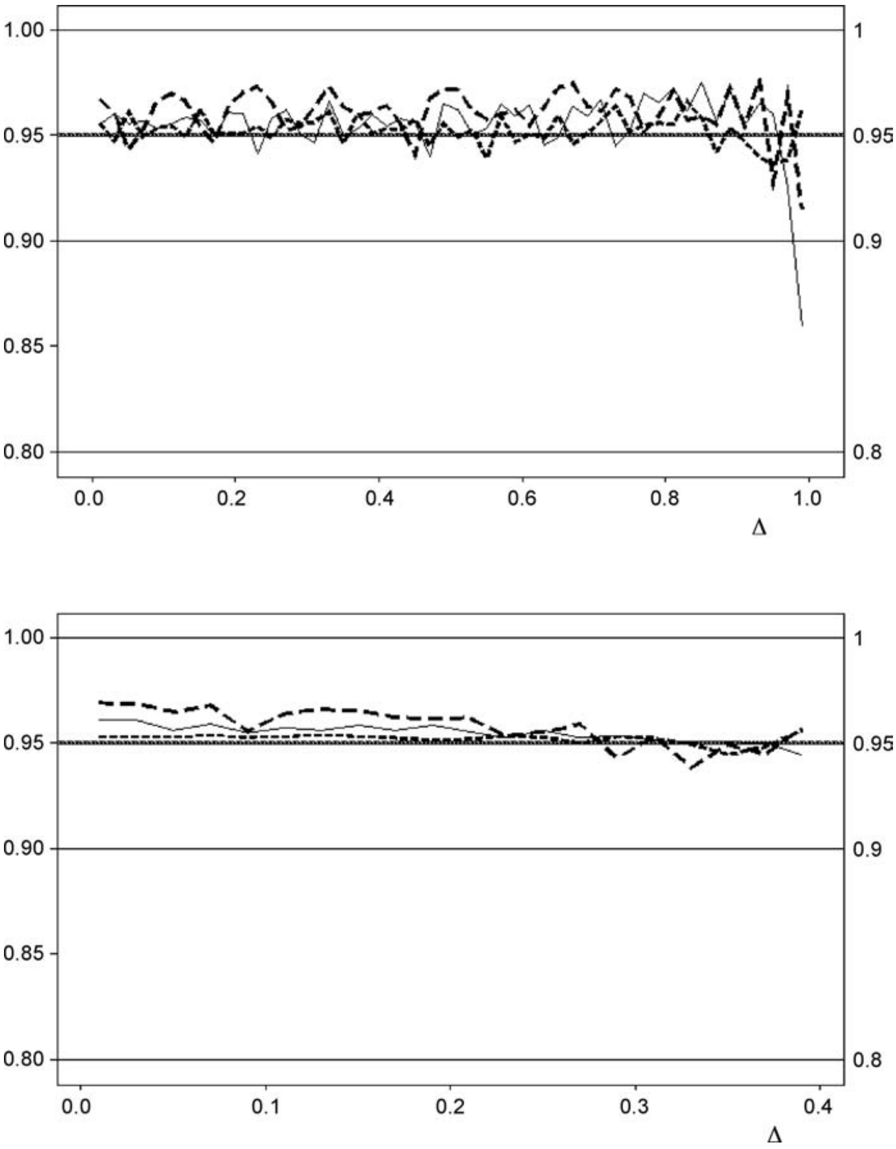


Figure 19.3 In the top plot are shown the empirical coverage probabilities of nominal 95 % confidence intervals for Δ based on (19.5) as a function of Δ. It is assumed $p = 0.5$. The graphs are shown for $n_1 = n_2 = 9$ as a dashed line, for $n_1 = n_2 = 15$ as a thin solid line and for $n_1 = n_2 = 30$ as a dotted line. In the lower plot $p = 0.2$ and sample sizes are again equal, with the graphs shown for $n_1 = n_2 = 15$ as a dashed line, for $n_1 = n_2 = 30$ as a thin solid line and for $n_1 = n_2 = 60$ as a dotted line.

clear that this S satisfies the properties of a measure of evidence, even for p_1, p_2 small, so we do not use it as such.

19.4.2 Two-sample odds ratio

The odds ratio of control to treatment is $\text{OR} = \{p_1/(1-p_1)\}/\{p_2/(1-p_2)\}$. One can estimate OR by substitution of $\tilde{p}_1$, $\tilde{p}_2$ for p_1, p_2 to obtain $\widetilde{\text{OR}}$. For small p_1, p_2 this estimator has a skewed distribution, so one again uses the log transformation. Let the log-odds ratio be defined by $\eta = \ln(\text{OR})$. As for the relative risk, standard asymptotics are employed to show $\tilde{\eta} = \ln(\widetilde{\text{OR}})$ has, for n_1, n_2 increasing without bound, a limiting normal distribution with asymptotic mean η and variance:

$$\text{Var}[\tilde{\eta}] = \frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}. \tag{19.7}$$

Thus the standard error of $\tilde{\eta}$ is $\text{SE}[\tilde{\eta}] = \left[1/\{n_1\tilde{p}_1(1-\tilde{p}_1)\} + 1/\{n_2\tilde{p}_2(1-\tilde{p}_2)\}\right]^{1/2}$.

One then obtains a $100(1-\alpha)$ % confidence interval $[L, U]$ for η by taking $L = \hat{\eta} - z_{1-\alpha/2}\,\text{SE}[\hat{\eta}]$ and similarly for U. This interval $[L, U]$ for η is transformed to a $100(1-\alpha)$ % confidence interval $[e^L, e^U]$ for OR.

One can standardize $\tilde{\eta}$ to obtain a statistic which is asymptotically normal with variance 1, but as with the log relative risk, it is not clear that this provides a measure of evidence for $\eta > 0$ over the parameter space.

19.4.3 New confidence intervals for the RR and OR

The confidence intervals derived for Δ in Section 19.1.1 are based on the evidence T, which resulted from a *vst* applied conditionally to the distribution of $\hat{\Delta}$, given $\tilde{p} = p$. In view of the fact that for fixed $p = qp_1 + (1-q)p_2$, the RR and OR are strictly increasing in Δ as shown in Section 19.1.4, the intervals for Δ can be transformed into intervals for the RR and OR, maintaining the same nominal conditional confidence coefficient. Further investigation into these intervals and comparison with the traditional intervals presented above are required before one can recommend them.

19.5 Recurrent urinary tract infections

Recurrent urinary tract infections are a common health problem, and treatment by different antibiotics at various dosages has been tested in a large number of case-control studies. A recent review by Albert *et al.* (2004) included a summary of 11 such studies. For background, references and standard meta-analytic results, the reader may consult the website at www.nicsl.com.au and follow the prompts. The data are listed in columns 2–5 of Table 19.1. Here x_1 is the number of n_1 control patients who had recurrent infections, while x_2 is the number of n_2 treated patients who had recurrent infections following treatment.

The quantities in columns 6–10 show for each study estimates of the risk difference $\tilde{\Delta}$, the unknown constants $\tilde{p}$, $\tilde{\zeta}$, the correlation effect size $\tilde{\rho}$ and evidence T for a reduction of risk of infection $\Delta > 0$. Definitions for each are given in Section 19.1.1.

Table 19.1 Results of 11 independent studies of antibiotic treatment to prevent recurrent urinary tract infection. For each study the number x_1 out of n_1 control subjects who continued to have infections is listed, as well as the number x_2 of n_2 of treated subjects. See text for details regarding results.

Study	x_1	n_1	x_2	n_2	$\tilde{\Delta}$	$\tilde{p}$	$\tilde{\zeta}$	$\tilde{\rho}$	T	L	U
1	17	22	1	23	0.698	0.419	0.975	0.707	5.27	0.47	0.88
2	17	19	8	21	0.489	0.643	0.921	0.509	3.38	0.22	0.75
3	4	13	2	15	0.165	0.245	0.743	0.192	1.02	−0.18	0.53
4	8	21	1	20	0.315	0.225	0.698	0.377	2.47	0.08	0.64
5	10	13	0	11	0.708	0.366	0.935	0.733	4.03	0.41	0.94
6	13	17	4	18	0.513	0.501	1.001	0.513	3.19	0.21	0.76
7	5	6	1	13	0.679	0.571	1.133	0.637	3.01	0.24	0.91
8	15	25	3	25	0.462	0.365	0.928	0.479	3.53	0.22	0.70
9	13	23	1	20	0.491	0.300	0.844	0.535	3.70	0.26	0.76
10	5	7	1	13	0.580	0.484	1.098	0.554	2.63	0.15	0.86
11	9	11	2	16	0.645	0.529	1.032	0.635	3.57	0.30	0.87

All estimated effects $\tilde{\Delta}$ are positive. The weighted average of risks $\tilde{p}$ vary widely from 0.192 to 0.708 and the estimates of correlation effect size $\tilde{\rho}$ are all positive but quite variable. The first study contains large evidence T for a positive effect due to treatment, and the third study very little evidence, but most show moderate evidence for a positive effect. These evidences are combined in Chapter 25.

The last two columns of Table 19.1 give confidence intervals $[L, U]$ for the risk difference Δ, with only the third study interval containing the null $\Delta = 0$.

19.6 Summary

For two samples, variance stabilization led to very good coverage properties of interval estimators of the risk difference $\Delta = p_1 - p_2$. Whether similar techniques can improve on traditional confidence intervals for the relative risk and odds ratio remains to be seen.

20

Defining evidence in *t*-statistics

20.1 Example

Mulrow *et al.* (2004) conducted a review of studies in which the drop in systolic blood pressure following a weight-reducing diet for a group of patients was compared to that of a control group. Here we only consider the results for the treated patients for three studies, but a comparison of treated with control groups for seven studies is given in Section 21.4.3. The data are summarized by sample size n, sample mean $\bar{y}_n$ and sample standard deviation s_n, shown in columns 2–4 of Table 20.1. Column 5 contains the Student t-statistic for each study denoted S_n, and column 6 the p-value.

20.2 Evidence in the Student *t*-statistic

Given n observations $Y_1, \ldots, Y_n$ from the normal model $N(\mu, \sigma^2)$, with both parameters unknown, we want a measure of the evidence against $\mu = \mu_0$ in favor of $\mu > \mu_0$. Define the effect by $\theta = \mu - \mu_0$ and the *standardized effect* by

$$\delta = \theta/\sigma = (\mu - \mu_0)/\sigma\,,$$

a ratio of two unknown parameters. Denote the sample mean and variance of the observations by $\bar{Y}_n$ and s_n^2. Recall from Chapter 1 that when $\sigma = \sigma_0$ is known, the evidence in $\bar{Y}_n$ for the one-sided alternative $\mu > \mu_0$ is defined by $T_0 = \sqrt{n}\,(\bar{Y}_n - \mu_0)/\sigma_0 \sim N(\tau, 1)$ with expected evidence $\tau = \sqrt{n}\,\theta/\sigma_0$. In this model with known variance, the key inferential function is thus equal to the standardized effect $\delta = \theta/\sigma_0$. We now want to define the evidence and expected evidence when σ is unknown.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

Table 20.1 One-sample data for each of three studies measuring drop in systolic blood pressure for treated patients undergoing a weight-loss regime.

Study	n	$\bar{y}_n$	s_n	S_n	p-value
1	27	−4.8	13.8	−1.81	8.2×10^{-2}
2	20	13.3	8.1	7.34	5.8×10^{-7}
3	66	11.0	17.1	5.23	1.9×10^{-6}

The t-statistic

$$S_n = \sqrt{n}\,(\bar{Y}_n - \mu_0)/s_n$$

has under the null hypothesis $\mu = \mu_0$ the Student t-distribution with $\nu = n - 1$ degrees of freedom, but in order to derive the evidence, we need to study its distribution under the alternative $\mu > \mu_0$. In this case, we can rewrite the t-statistic as

$$S_n = \frac{\sqrt{n}\,(\bar{Y}_n - \mu) + \sqrt{n}\,(\mu - \mu_0)}{s_n},$$

which is known to have the noncentral t-distribution.

Definition 20.1 *The random variable X defined as a function of two independent random variables $Z \sim \mathcal{N}(0,1)$ and $W \sim X_\nu^2$ by*

$$X = \frac{Z + \lambda}{\sqrt{W/\nu}}$$

is said to have a noncentral Student's $t_\nu(\lambda)$ distribution. The noncentrality parameter $\lambda \in \mathbb{R}$ and the number of degrees of freedom $\nu \in \{1, 2, 3, \ldots\}$ are the two parameters that characterize this law.

When $\mu > \mu_0$, the t-statistic S_n has a noncentral t-distribution with parameters $\nu = n - 1$ and $\lambda = \sqrt{n}\,\delta$. This follows by putting $Z = \sqrt{n}(\bar{Y}_n - \mu)/\sigma$ and $W = s_n^2/\sigma^2$. A good reference for the noncentral t-distribution is Chapter 31 of Johnson *et al.* (1995).

Letting $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{e}^{-x} \mathrm{d}x$ denote the gamma function of $\alpha > 0$, the first two moments of the noncentral t-distribution $X \sim t_\nu(\lambda)$ are

$$\begin{aligned} \mathrm{E}[X] &= c_\nu\,\lambda = \left(\frac{\nu}{2}\right)^{1/2} \frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}\;\lambda \\ &= \left(1 + \frac{3}{4\nu} + O(1/\nu^2)\right)\lambda \qquad (20.1) \\ \mathrm{E}[X^2] &= \frac{\nu}{\nu - 2}\,\{1 + \lambda^2\} \end{aligned}$$

$$var[X] = \frac{\nu}{\nu - 2} + \lambda^2 \left(\frac{\nu}{\nu - 2} - c_\nu^2 \right)$$
$$= 1 + \frac{\lambda^2}{2\nu} + O(1/\nu^2) = 1 + \frac{\mathrm{E}[X]^2}{2\nu} + O(1/\nu^2). \tag{20.2}$$

The expansions into powers of $1/\nu$ are valid when $\nu \to \infty$ and are based on Stirling's approximation for the gamma function. Using these approximations, one finds that the *vst* appropriate when $X \sim t_\nu(\lambda)$ has to satisfy

$$(h'(x))^2\,(1 + x^2/(2\nu)) = 1$$

for all x (see Section 17.2). The solution is

$$h(x) = \sqrt{2\nu}\ \sinh^{-1}\left(\frac{x}{\sqrt{2\nu}} \right). \tag{20.3}$$

The sinh function, an abbreviation for *hyperbolic sine function* is defined by $\sinh(x) = (\mathrm{e}^x - \mathrm{e}^{-x})/2$ for all x. It has inverse function $\sinh^{-1}(x) = \ln(x + \sqrt{x^2 + 1})$ for all x. The reader can verify that the inverse function is an odd function ($\sinh^{-1}(-x) = -\sinh^{-1}(x)$), and that its first derivative is $\frac{\mathrm{d}}{\mathrm{d}x}\sinh^{-1}(x) = \{1 + x^2\}^{-1/2}$. Equation (20.3) is due to Azorin (1953), who first studied variance stabilization of the noncentral t-distribution and derived a more elaborate and more accurate formula. To define the evidence in the t-statistic, we apply the *vst* (20.3) to the t-statistic S_n, but introduce a further simplification. In the case of the t-statistic, $\nu = n - 1$. Thus,

$$h(S_n) = \sqrt{2(n-1)}\ \sinh^{-1}\left(\frac{S_n}{\sqrt{2(n-1)}} \right).$$

Because S_n contains a factor of $\sqrt{n}$, we could further simplify the expression, if we used $\nu = n$ instead of $\nu = n - 1$. This introduces a small error of order $O(1/n)$, which is of the same order as the other simplifications we made.

Definition 20.2 *The evidence in a t-statistic for testing $\mu = \mu_0$ against the alternatives $\mu > \mu_0$ is*

$$T = \sqrt{2n}\ \sinh^{-1}\left(\frac{S_n}{\sqrt{2n}} \right) = \sqrt{2n}\ \sinh^{-1}\left(\frac{(\bar{Y}_n - \mu_0)/s_n}{\sqrt{2}} \right). \tag{20.4}$$

As we will see in Section 20.4 the finite sample corrected evidence

$$T_{\text{corrected}} = \left(\frac{n - 1.7}{n - 1} \right) \sqrt{2n}\, \sinh^{-1}(S_n/\sqrt{2n}). \tag{20.5}$$

improves the normal approximation in the tails.

For the data of Section 20.1, the evidence is shown in Table 20.2.

Let θ_k be the unknown drop in systolic blood pressure for the kth treated group. Then for the first study $T_1 = -1.79$ is weak evidence for the alternative $\theta_k < 0$ to the null hypothesis of no effect. However, $T_2 = 6.26$ and $T_3 = 5.06$ provide strong evidence for the respective alternatives $\theta_2 > 0$ and $\theta_3 > 0$. All the T_k's have standard error 1. The main question is how to combine these evidence values, and that is the topic for later chapters. But first, in this chapter we justify the choice of evidence measure T.

Table 20.2 One-sample data for each of three studies measuring drop in systolic blood pressure for treated patients undergoing a weight-loss regime. Instead of the *p*-value, the evidence is indicated.

Study	n	$\bar{y}_n$	s_n	S_n	Evidence T
1	27	−4.8	13.8	−1.81	−1.79
2	20	13.3	8.1	7.34	6.26
3	66	11.0	17.1	5.23	5.06

Laubscher (1960) has studied the question whether or not the evidence is normally distributed and found problems for small values of n and large values of the noncentrality parameter. If the normal approximation for the evidence were to hold, we would have $T \sim N(\sqrt{n}(\sqrt{2}\sinh^{-1}(\delta/\sqrt{2})), 1)$. The quality of this approximation can be checked in various ways, for example, by comparing quantiles, distribution functions or densities and we will come back to this in more detail a little later on. For now, let us simply look at how well we do based on *p*-values. Figure 20.1 shows the *p*-value after the application of the *vst* as a function of the *p*-value before the transformation to evidence. The *p*-value often is based on the normal approximation, the *p*-value before is computed with the central *t*-distribution; see Section 20.4.1.

We can also check the approximation by looking at densities. The actual cumulative distribution of the evidence T in (20.4) is

$$\begin{aligned} P(T \leq t) &= P\left(\sqrt{2n}\sinh^{-1}\left(\frac{S_n}{\sqrt{2n}}\right) \leq t\right) \\ &= P\left(S_n \leq \sqrt{2n}\sinh\left(\frac{t}{\sqrt{2n}}\right)\right) \\ &= F_{\text{Student}}\left(\sqrt{2n}\sinh\left(\frac{t}{\sqrt{2n}}\right)\middle| n-1, \sqrt{n}\delta\right), \end{aligned}$$

where $F_{\text{Student}}(t|v, \lambda)$ is the cumulative distribution function of the noncentral *t*-distribution with v degrees of freedom and noncentrality λ. From this formula the density is easy to derive.

Figure 20.2 shows the densities in six representative situations together with the approximate normal density. The three panels in the top row are for $n = 5$, the three in the bottom row are for $n = 10$. The two left-most panels show the case of the central *t*-density. The actual density of the evidence has a variance that is slightly larger then one, the variance the approximating normal density. This effect is quite visible when $n = 5$, but much less pronounced when $n = 10$. The two right-most panels have a noncentrality of $\lambda = 4$. In these cases, the actual density of the evidence remains visibly asymmetric and the density of the evidence T again has a variance slightly bigger than one. In addition, the mode of the approximating normal density is to the left of the mode of the actual density. Again, we also note that these effects diminish with increasing n. The two middle panels are for $\lambda = 2$.

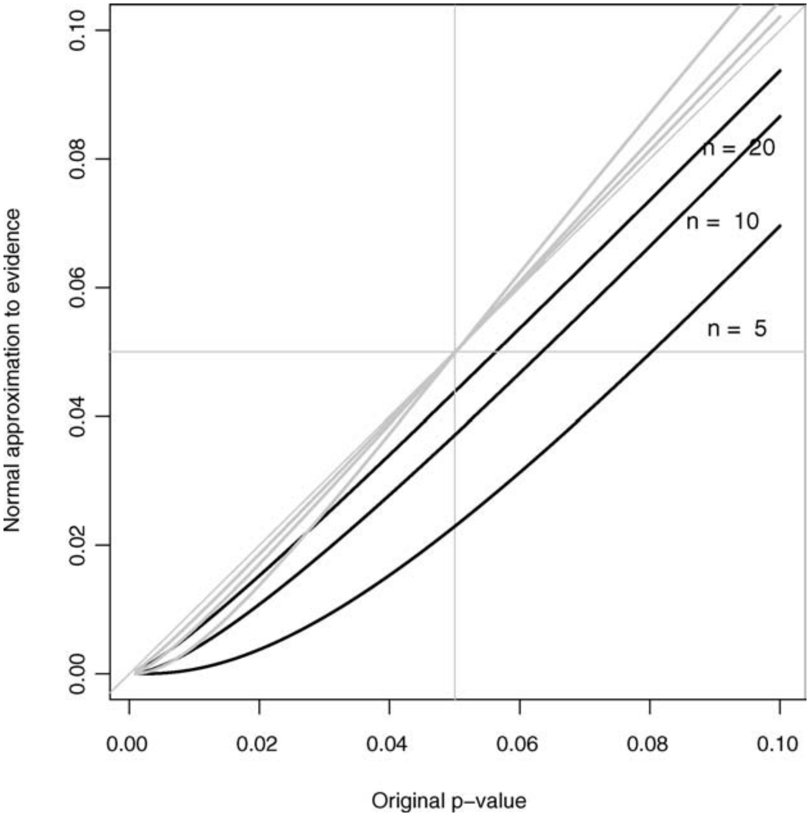


Figure 20.1 The curves show the p-values before and after using the *vst*, for different sample sizes. Ideally, all the curves would lie exactly on the diagonal of the chart, which is clearly not the case here. After our transformation (20.4), the p-values based on the evidence are systematically too small. The curves in gray correspond to the corrected evidence and here the approximation is good.

20.3 The Key Inferential Function for Student's model

The expected evidence E[T] is approximately equal to

$$\begin{aligned}\mathrm{E}[T] &\doteq \sqrt{2n}\ \sinh^{-1}(\sqrt{n}\,\delta/\sqrt{2n}) \\ &= \sqrt{n}(\sqrt{2}\sinh^{-1}(\delta/\sqrt{2})) = \sqrt{n}\,\mathcal{K}(\delta)\,,\end{aligned} \tag{20.6}$$

where we made use of the approximate expectation of the t-statistic, which is $\sqrt{n}\,\delta$. Of course, we want the evidence T based on n observations to satisfy $T \sim N(\tau, 1)$, where $\tau = \sqrt{n}\,\mathcal{K}(\delta)$ for some monotonically increasing function of $\delta = (\mu - \mu_0)/\sigma$, called the Key Inferential Function. And this is exactly what seems to happen.

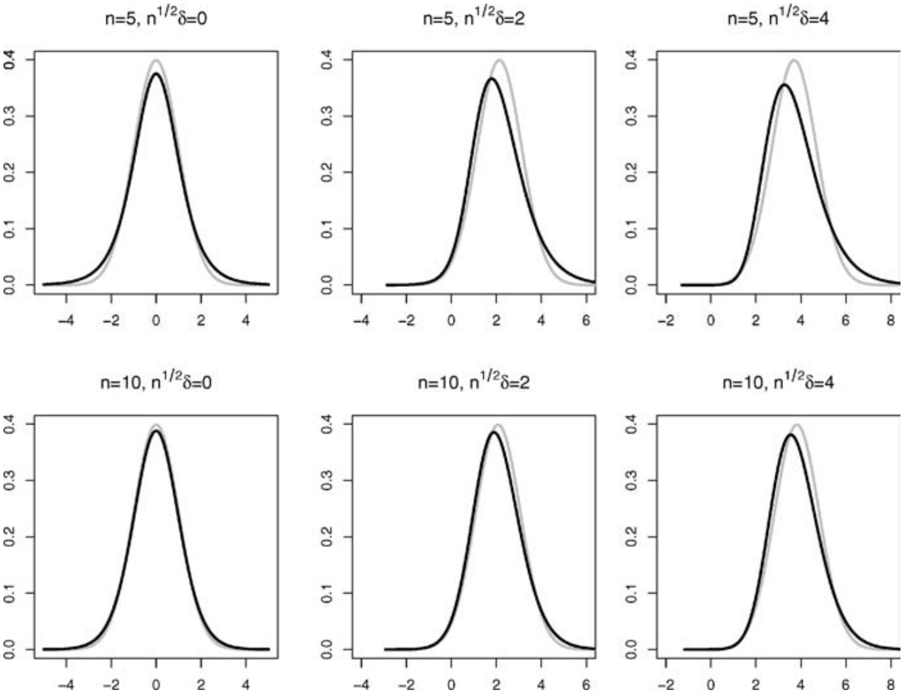


Figure 20.2 The six panels show the density of the evidence T (20.4) in situations with small sample sizes. The curves in gray are the approximating normal densities.

Definition 20.3 *The Key Inferential Function appropriate for Student's t-test is defined in (20.6) and equals*

$$\mathcal{K}(\delta) = \sqrt{2}\,\sinh^{-1}(\delta/\sqrt{2}) = \sqrt{2}\,\ln(\delta/\sqrt{2} + \sqrt{1+\delta^2/2}). \qquad (20.7)$$

The graph is plotted in Figure 20.3.

The Key Inferential Function transforms the standardized effect δ into *transformed standardized effect* $\mathcal{K}(\delta)$. The latter is estimated by $\hat{\kappa} = T/\sqrt{n}$.

To achieve 'moderate' expected evidence 3.3 for the one-sided alternative when $\delta = 0.5$ one would need a sample size of n satisfying $\sqrt{n}\,\mathcal{K}(0.5) = 3.3$, or $n = 35$. The evidence as usual has standard error 1 in estimating this expected value. We now can compare the expected evidence for $\mu > 0$ when σ is known with the expected evidence when σ is unknown. The former is $\sqrt{n}\,\delta$, while the latter is $\sqrt{n}\,\mathcal{K}(\delta)$. The ratio is equal to $\{\sinh^{-1}(\delta/\sqrt{2})\}/(\delta/\sqrt{2})$, which is approximately 1 for small δ but behaves like $\{1+\delta^2/2\}^{-1/2}$ for large δ.

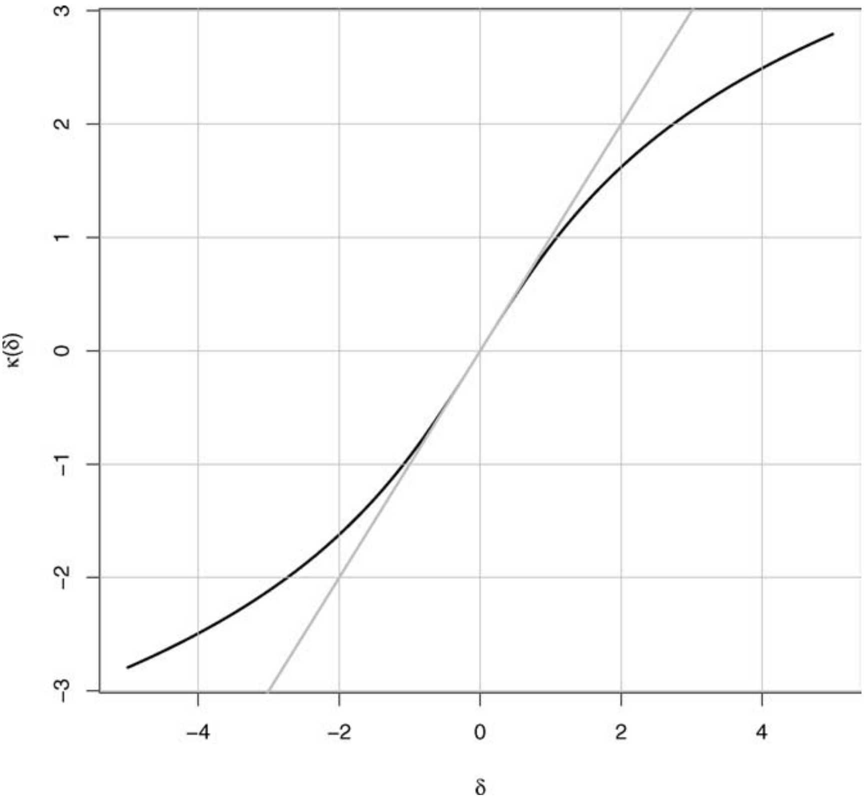


Figure 20.3 The Key Inferential Function for Student's model is shown as a function of the standardized effect δ.

Or, one can ask how much more work is required to obtain the same amount of expected evidence. If m observations with σ known and n observations with σ unknown are to lead to the same expected evidence, then the ratio of n to m satisfies $\sqrt{n/m} \doteq 1 + \delta/6$ over the range $1 < \delta < 20$, so the extra work required relative to m is $(n - m)/m \doteq (1 + \delta/6)^2 - 1 = \delta/3 + \delta^2/36$. The extra work required, because σ is unknown, is a quadratic function of the unknown standardized effect δ over most of the range of interest. In particular for $\delta = 3$, n must be roughly twice as large as m to achieve the same expected evidence. When $\delta = 6$, $n = 4m$ is required. We will come back to this question in Section 20.6.

20.4 Corrected evidence

The expected value of the evidence satisfies $\mathrm{E}[T] \doteq \sqrt{n}\,\mathcal{K}(\delta)$. In this section we discuss possible modifications of the evidence T with the aim of improving the precision of the above approximate equality. Recall that $\doteq$ means equal up to some wiggle room. We use this wiggle room to simplify the choice of expressions. Both our key $\mathcal{K}$ (20.7) and our evidence T (20.4) are easy-to-use formulas. But the imprecision in $\doteq$ may lead to appreciable errors for small samples. Thus the desire to correct the evidence.

20.4.1 Matching *p*-values

One of the ways in which the approximation error is made visible is in the *p*-value. Suppose we had a study and we wished to test $\mu = \mu_0$ against $\mu > \mu_0$. The result of the study can be summarized by its *p*-value p_{study} and the sample size n. We know how the *p*-value was calculated. It is equal to

$$p_{\text{study}} = F_S(-\sqrt{n}\,(\bar{Y}_n - \mu_0)/s_n) = F_S(-S_n),$$

where F_S denotes the cumulative distribution function of a *t*-distribution with $n-1$ degrees of freedom. We transform this information to evidence by way of

$$T_n = h_n(S_n) = \sqrt{2n}\,\sinh^{-1}(S_n/\sqrt{2n})$$

and could derive a *p*-value for the evidence as

$$p_{\text{evidence}} = \Phi(-T_n).$$

By construction, it is evident that $p_{\text{study}} \doteq p_{\text{evidence}}$. However, exact equality does not hold, which may to some users be disconcerting, especially when the study *p*-value is below 5 %, but the evidence *p*-value is above this bound. This particular problem is, of course, caused by the mistaken impression that the limit of 5 % is somehow sacred. So, our primary response to this worry is to emphasize that one should not rely on the *p*-value in the first place. It is much easier to think in terms of evidence. Because evidence has a standard error of 1, whether the *p*-value is slightly below 5 % or slightly above 5 % makes no real difference.

If, however, the translation to evidence has to preserve the *p*-value, we suggest using a correction. To have equality between the two requires

$$\Phi(-T_n) = F_S(-S_n) \Leftrightarrow T_n = h_n(S_n) = -\Phi^{-1}\left(F_S(-S_n)\right).$$

A multiplicative correction to the *vst* of the form

$$\tilde{h}_n(x) = (1 + c_n)\,h_n(x),$$

with c_n of order $O(1/n)$ would be appropriate. It amounts to a finite sample correction of lower order than the terms deleted in 20.1. The correction corresponds simply to a multiplication of the evidence T and Figure 20.2 can guide us in the choice of c_n. If we multiplied the evidence T by a constant slightly smaller than one, both deficiencies we noted in the discussion of these densities could be corrected, the mode of the density would shrink towards zero and the variance would become smaller. To choose a particular value of c_n, we could, for example, demand that a study *p*-value of 5%

translate into an evidence *p*-value of 5%. Using the symmetry of the *t*-distribution F_S, this implies that

$$\begin{aligned}\Phi(-\tilde{h}_n(F_S^{-1}(0.95)) &= F_S(-F_S^{-1}(0.95)) = F_S(F_S^{-1}(0.05)\\ (1+c_n)\,{}_n(F_S^{-1}(0.95)) &= -\Phi^{-1}(0.05) = \Phi^{-1}(0.95)\\ c_n &= \Phi^{-1}(0.95)/h_n(F_S^{-1}(0.95)) - 1.\end{aligned}$$

This correction anchors the *vst* at the 95 % quantiles of the Student *t*-distribution and the normal distribution. The correcting constant c_n turns out to be nearly equal to

$$c_n \doteq -0.7/(n-1).$$

The corrected *vst* thus leads to the following evidence

$$T_{\text{corrected}} = \left(1 - \frac{0.7}{n-1}\right)\sqrt{2n}\,\sinh^{-1}(S_n/\sqrt{2n})\,.$$

The fact that the correction has a negative sign means that the corrected evidence is smaller than the uncorrected one. The corrected evidence thus has a bigger associated *p*-value $\Phi(-T_{\text{corrected}})$.

Mathematical statistics offers some additional insight to the problem of matching *p*-values. The Cornish–Fisher expansion is a tool for matching all the quantiles of an arbitrary distribution to those of a normal distribution. For the *t*-distribution it leads to

$$F_S^{-1}(p) \doteq \Phi^{-1}(p) + [(\Phi^{-1}(p))^3 - 3\Phi^{-1}(p)]\,\frac{1}{8n}\,.$$

This formula could be used to derive a more general expression for the correction c_n.

20.4.2 Reducing bias

Another basis for choosing a correction is the bias incurred when estimating the expected evidence $E(T) = \tau \doteq \sqrt{n}\mathcal{K}(\delta)$. The plug-in estimate of $E(T)$ is simply the observed value $\hat{\tau} = T$. This is equivalent to the use of

$$\hat{\kappa} = \mathcal{K}(\hat{\delta}_n) = \mathcal{K}\left(S_n/\sqrt{n}\right)$$

based on the straightforward estimate of the standardized effect, $\hat{\delta}_n = (\bar{Y}_n - \mu_0)/s_n = S_n/\sqrt{n}$. In order to realize the full potential of the Key Inferential Function, we need an unbiased estimator $\hat{\mathcal{K}}_{\text{corrected}}$ for which $\sqrt{n}\,\hat{\mathcal{K}}_{\text{corrected}} \sim N(\sqrt{n}\,\mathcal{K}(\delta), 1)$. This will enable us to find confidence intervals for δ and to combine evidence from different but related experiments, as described in Chapters 17 and 25.

There are two distinct sources of bias in the basic estimate $\hat{\mathcal{K}}$. First, as shown in (20.1), $\hat{\delta}_n$ is not an unbiased estimate of δ, and, second, even if it were, $\hat{\mathcal{K}}$ would still be biased, because $\mathcal{K}$ is nonlinear. To compute the bias, consider the three-term Taylor expansion

$$\mathcal{K}(\hat{\delta}_n) \doteq \mathcal{K}(\delta) + (\hat{\delta}_n - \delta)\,\mathcal{K}'(\delta) + (\hat{\delta}_n - \delta)^2\,\mathcal{K}''(\delta)/2.$$

Taking the expectation leads to

$$\mathrm{E}[\mathcal{K}(\hat{\delta}_n)] \doteq \mathcal{K}(\delta) + (\mathrm{E}[\hat{\delta}_n] - \delta)\,\mathcal{K}'(\delta) + \mathrm{E}[(\hat{\delta}_n - \delta)^2]\,\mathcal{K}''(\delta)/2.$$

From (20.1) we find

$$\mathrm{E}[\hat{\delta}_n] \doteq \delta(1 + 3/(4n)) \text{ and } \mathrm{E}[(\hat{\delta}_n - \delta)^2] \doteq (1 + \delta^2/2)/n.$$

Putting these two together yields

$$\mathrm{E}[\mathcal{K}(\hat{\delta}_n)] \doteq \mathcal{K}(\delta) + 3\delta\,\mathcal{K}'(\delta)/(4n) + (1 + \delta^2/2)\,\mathcal{K}''(\delta)/(2n).$$

We leave it to the reader to check that $\mathcal{K}'(\delta) = (1+\delta^2/2)^{-1/2}$ and $\mathcal{K}''(\delta) = -\delta\,(\mathcal{K}'(\delta))^3/2$. The above result thus simplifies to

$$\mathrm{E}[\mathcal{K}(\hat{\delta}_n)] \doteq \mathcal{K}(\delta) + 3\delta\,\mathcal{K}'(\delta)/(4n) - \delta\,\mathcal{K}'(\delta)/(4n) \doteq \mathcal{K}(\delta(1 + 1/(2n))).$$

When matching *p*-values we found that the uncorrected estimate (20.4) slightly overestimates the evidence. This is confirmed by the above computation. The estimate $\hat{\kappa}$ overestimates the Key Inferential Function by an amount that is inversely proportional to the size of the study. This suggests a correction of $\hat{\kappa}$, namely

$$\hat{\mathcal{K}}_{\text{unbiased}} = \mathcal{K}(\hat{\delta}_n[1 - 1/(2n)]) = \sqrt{2}\,\sinh^{-1}(\hat{\delta}_n(2n - 1)/(2n)). \tag{20.8}$$

In order to compute the evidence, $\mathcal{K}$ must be multiplied by $\sqrt{n}$. This gives an alternative bias-corrected *vst*, namely

$$T_{\text{unbiased}} = \sqrt{2n}\,\sinh^{-1}((2n - 1)S_n/(2n\sqrt{2n})). \tag{20.9}$$

For the data of Section 20.1, the three versions of the evidence are shown in Table 20.3.

Figure 20.4 demonstrates that the bias correction does indeed provide an improvement on (20.4) by reducing the bias. The variance stabilization is, however, much better with the finite sample correction, where the standard error is very nearly equal to one, even for samples of size 10. For this reason, we do not recommend the use of the evidence T_{unbiased} in practice.

Table 20.3 One-sample data for each of three studies measuring drop in systolic blood pressure for treated patients undergoing a weight-loss regime. The three versions of evidence are shown in the last three columns. The corrections are of small size, but show that the raw evidence overstates the true value somewhat.

Study	n	S_n	Raw	Corrected	Unbiased
1	27	−1.81	−1.79	−1.74	−1.76
2	20	7.34	6.26	6.03	6.14
3	66	5.23	5.06	5.01	5.03

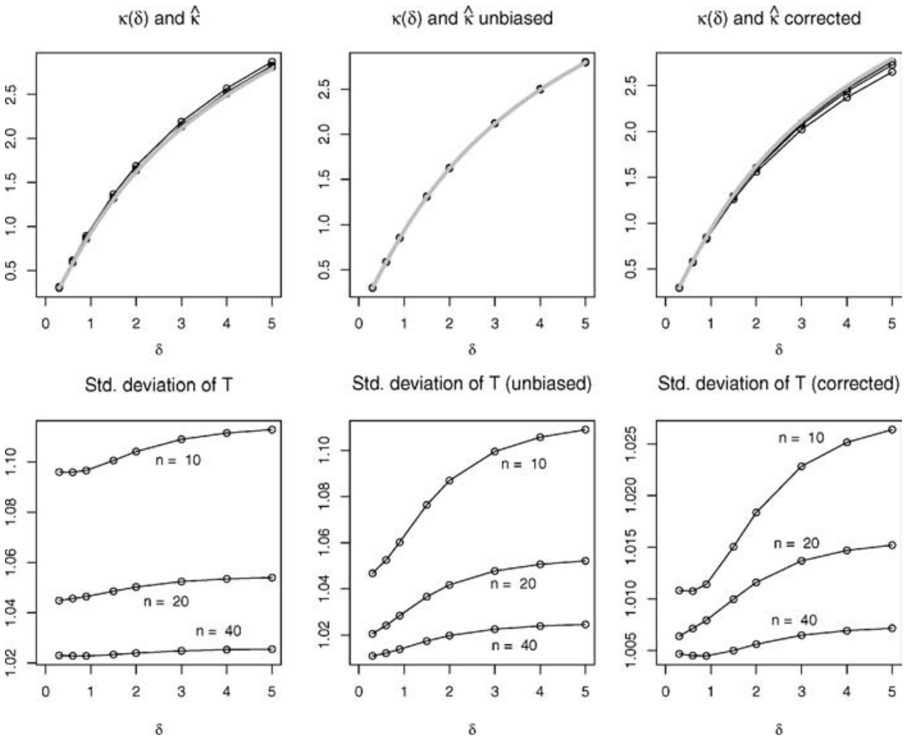


Figure 20.4 In the top row of panels, the value of $\mathcal{K}(\delta) = \sqrt{2}\sinh^{-1}(\delta/\sqrt{2})$ is shown as a thick gray line. Also shown are the expected values of the estimators $\hat{\kappa}$, $\hat{\kappa}_{\text{corrected}}$ and $\hat{\kappa}_{\text{unbiased}}$ for study sizes of $n = 10, 20, 40$. The expected values were computed by taking the mean over 100,000 simulations. The panels in the bottom row show the corresponding estimated standard deviations of the evidences T, $T_{\text{corrected}}$, and T_{unbiased}. The ordinate of each plot is provided by the standardized effect δ.

20.5 A confidence interval for the standardized effect

How to construct a confidence interval for the standardized effect $\delta = (\mu - \mu_0)/\sigma$ is the topic of this section.

The estimated key inferential statistic $\hat{\kappa}$ is approximately normal with mean $\mathcal{K}(\delta) = \sqrt{2}\,\sinh^{-1}(\delta/\sqrt{2}\,)$ and variance $1/n$ for a range of values of δ. A nominal 95 % confidence interval for $\mathcal{K}(\delta)$ is given by $[\hat{\kappa} \pm 1.96/\sqrt{n}]$. Because $\mathcal{K}(\cdot)$ is a one-to-one function, this can easily be inverted to yield a confidence interval for the standardized effect δ. The inverse function is $\delta = \sqrt{2}\,\sinh(\mathcal{K}/\sqrt{2}\,) = \{e^{\mathcal{K}/\sqrt{2}} - e^{-\mathcal{K}/\sqrt{2}}\}/\sqrt{2}$, which leads to a nominal 95 % confidence interval for δ in terms of the evidence $T = \sqrt{n}\,\hat{\mathcal{K}}$ of

$$[L, U] = \left[\sqrt{2}\,\sinh\left(\frac{T - 1.96}{\sqrt{2n}}\right), \sqrt{2}\,\sinh\left(\frac{T + 1.96}{\sqrt{2n}}\right)\right]. \tag{20.10}$$

As seen above, the corrected evidence T_{unbiased} is preferable and for smallest sample sizes. When using the corrected evidence (20.5), the inversion formula leads to the confidence interval

$$[L_{\text{corrected}}, U_{\text{corrected}}] = \left[\sqrt{2}\,\sinh\left(\frac{T_{\text{corrected}} - 1.96}{\frac{n-1.7}{n-1}\sqrt{2n}}\right), \sqrt{2}\,\sinh\left(\frac{T_{\text{corrected}} + 1.96}{\frac{n-1.7}{n-1}\sqrt{2n}}\right)\right],$$

which is always a bit wider than the confidence interval (20.10).

For the data of Section 20.1, the sample sizes are so large that the correction is not necessary. The first sample based on $n = 27$ observations has $S_{27} = -1.81$ which leads to $\hat{\delta} = -0.336$ and a 95 % confidence interval for δ from (20.10) of $[L, U] = [-0.75, 0.03]$. For the next sample $S_{20} = 7.3$, $\hat{\delta} = 1.35$ and $[L, U] = [1.03, 2.40]$, and the third $S_{66} = 5.23$, $\hat{\delta} = 0.62$ and $[L, U] = [0.39, 0.92]$.

20.5.1 Simulation study of coverage probabilities

In Figure 20.5 are shown the empirical coverage probabilities based on 100 000 simulations each of nominal 95 % confidence intervals $[L, U]$ for δ, when (20.10) is employed with $T_{\text{unbiased}} = \sqrt{n}\,\hat{\mathcal{K}}_{\text{unbiased}}$ based on the bias-corrected *vst* given by (20.8).

20.6 Comparing evidence in *t*- and *z*-tests

20.6.1 On substituting *s* for *σ* in large samples

The expected evidence in a one-sided t-test of $\mu = \mu_0$ against $\mu > \mu_0$ is equal to $\sqrt{n}\mathcal{K}(\delta)$, whereas for the z-test it is equal to $\sqrt{n}\delta$. For large $\delta = (\mu - \mu_0)/\sigma$ there is thus a very notable difference in the evidence obtained from a t-statistic $S = \sqrt{n}(\bar{Y}_n - \mu_0)/s_n$ compared to a z-test $Z = \sqrt{n}(\bar{Y}_n - \mu_0)/\sigma$. Clearly, as the sample size n grows the standard deviation becomes more or less ‘known’, the distributions of the two test statistics under the null hypothesis are ‘equal’ and the decisions taken for or against the null hypothesis of the two tests are in agreement. The reason for the difference in the evidences lies in the different behavior of the distribution of the two test statistics when the alternative hypothesis holds. In Section 20.2 it was shown that the distribution of the t-test has a noncentral t-distribution $t_\nu(\lambda)$ with $\nu = n - 1$ and $\lambda = \lambda(\nu) = \sqrt{n}\,\delta = \sqrt{(\nu + 1)}\,\delta$. If we let the sample size grow, the distribution of $X \sim t_\nu(\lambda(\nu))$ is for large ν delicately poised between normality and a skewed distribution. The following proposition explains in detail what happens.

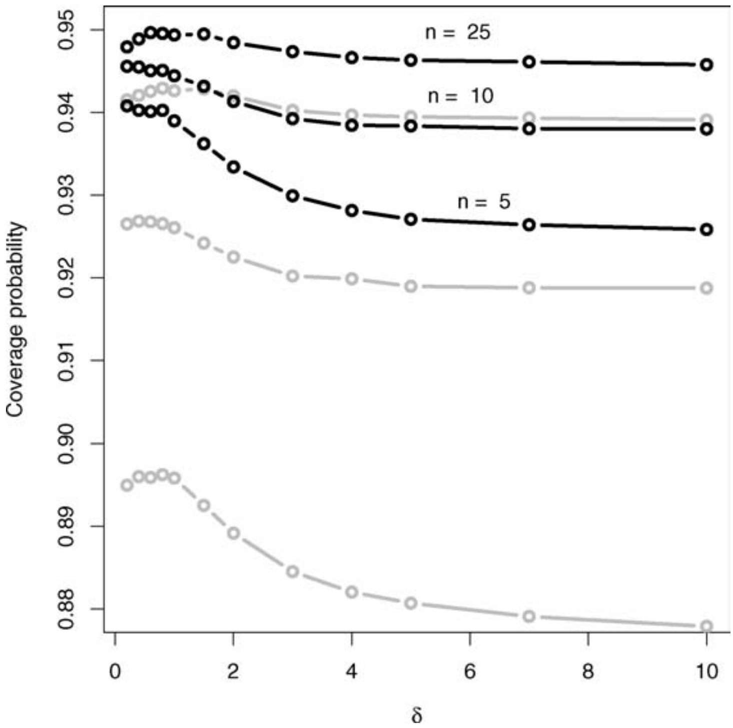


Figure 20.5 The coverage probabilities for the standardized effect δ are plotted as functions of δ for sample sizes $n = 5, 10, 25$. The curves in grey are for the evidence T (20.4), the black curves are for the corrected evidence $T_{\text{corrected}}$ (20.5). The corrected evidence leads to coverage probabilities close to 95% over a wide range of situations.

Proposition 20.1 *Let $X \sim t_\nu(\lambda(\nu))$. There are three cases of possible convergence as $\nu \to \infty$, depending on the rate of growth of $\lambda(\nu)$ with ν:*

$$\begin{cases} \textit{if } \lambda(\nu) \to \lambda, & \textit{then } X \to N(\lambda, 1); \\ \textit{if } \lambda(\nu) = \sqrt{\nu}\,\delta, & \textit{then } X - \sqrt{\nu}\,\delta \to N(0, 1 + \delta^2/2); \\ \textit{if } \lambda(\nu) = \nu^k\,\lambda \textit{ for } k > 1/2, & \textit{then } (X - \lambda(\nu))/\nu^{k-1/2} \to N(0, \delta^2/2). \end{cases}$$

Proof: By definition (20.1), we can write X as

$$\frac{Z}{\sqrt{W/\nu}} + \frac{\lambda(\nu)}{\sqrt{W/\nu}},$$

where Z is a standard normal random variable and W is independent of Z and has a χ^2_ν distribution. The mean and variance of W are ν and 2ν. The mean and variance of $1/\sqrt{W/\nu}$ are thus approximately equal to 1 and $2\nu/(2\nu)^2 = 1/(2\nu)$.

As $\nu \to \infty$, the first summand converges to a normal limit, and the standardized variable $\sqrt{2\nu}\,(1/\sqrt{W/\nu} - 1)$ also tends to a normal distribution. In other words,

$\lambda(\nu)/\sqrt{W/\nu}$ is approximately normal with mean $\lambda(\nu)$ and variance $\lambda(\nu)^2/(2\nu)$. The three different cases correspond to different behaviors of the second summand.

In the first case, $\lambda(\nu) \to \lambda$ as $\nu \to \infty$, the second term tends in probability to λ and the sum of the two converges to a normal with mean λ and variance 1. This situation occurs when performing a t-test with alternatives that depend on the sample size n and approach the null hypothesis as n grows. These so-called contiguous alternatives are of the form $\mu_0 + \theta/\sqrt{\nu}$. For these alternatives the fact that the standard deviation σ has to be estimated has no importance and the two tests are equivalent.

The second case occurs when we consider the t-test under a growing sample size but for a fixed alternative $\mu > \mu_0$. Here, the second summand is approximately normal with mean $\sqrt{\nu}\,\delta$ and variance $\delta^2/2$. Since the two summands are independent, their sum is approximately normal with mean $\sqrt{\nu}\,\delta$ and variance $1 + \delta^2/2$. In this case, the t-test and the z-test are not equivalent. The t-test has a bigger variance due to the fact that the standard error needs to be estimated.

In the third case, $\lambda(\nu) = \nu^k\,\delta$ as $\nu \to \infty$, the second summand has an approximate normal distribution with mean $\nu^k\,\delta$ and variance $\nu^{2k-1}\delta^2/2$. Here we are interested in $2k - 1 > 0$, in which case the variance grows with ν and we need to divide by $\nu^{k-1/2}$ in order to standardize. Doing this causes the first summand to converge to zero in probability, thus proving the stated limit. This corresponds to a situation where with growing sample size n the alternatives move away from the null hypothesis. In this case it is the variance of the estimate of the standard deviation which determines the limit.

Except in the first case of contiguous alternatives, there is thus a real difference between the z-test and the t-test. This difference shows itself in a reduction of the power of the t-test. One has to note, however, that the practical relevance of this phenomenon is minor. The difference in the expected evidence is $\sqrt{n}\,(\delta - \mathcal{K}(\delta))$. This is going to be large in absolute value only if either the sample size n or the standardized effect δ or both are large. In either case the expected evidence carried by a t-statistic is large as well and the power is close to one. The contiguous alternatives have originally been invented exactly as a response to this conundrum. If one makes experiments with large n then the power will be close to one even for small effects δ and, in order to study the power more closely, contiguous alternatives have to be used.

20.7 Summary

For normal data the test statistic for $\mu = \mu_0$ is the Student t-statistic, and its variance stabilization was essentially solved by Azorin (1953). However, we require a little more, because when we combine evidence from different studies in later chapters, it will be important that the *expected* evidence in each be of the form $\sqrt{n}\,\mathcal{K}(\delta)$ where $\mathcal{K}$ is monotonically increasing in the unknown standardized effect $\delta = (\mu - \mu_0)/\sigma$, and that $\mathcal{K}$ be the same for all studies. Therefore we reexpressed Azorin's *vst* in terms which enabled us to obtain unbiased estimates $\hat{\kappa}_k$ from different studies for which each $T_k = \sqrt{n_k}\,\hat{\kappa}_k \sim N(\sqrt{n_k}\,\mathcal{K}(\delta), 1)$.

It turns out that the appropriate $\mathcal{K}(\delta) = \sqrt{2}\ \sinh^{-1}(\delta/\sqrt{2})$. Knowing this function determines the expected evidence in the Student t-statistic enables us to assess what to expect for a given sample size n and standardized effect δ. It also allows us to choose the sample size to obtain a desired amount of expected evidence for a particular alternative δ.

In order to estimate $\sqrt{n}\,\mathcal{K}(\delta)$ with standard error 1 we introduced two different but similar bias-corrections. The benefits of finding a good *vst* become apparent when calculating confidence intervals for the standardized effect δ, where for all practical purposes a sample size of 10 is sufficient to attain the nominal coverage of 95 % for δ ranging from -2 to 2. This range includes what Cohen (1988) considers a 'large' effect, $\delta = 0.8$, and what W.G. Hopkin (http://sportsci.org/resource/stats/) calls 'very large'. The range of δ for which this method yields accurate confidence intervals increases with increasing sample size.

Next we examined the difference between the t-test and the z-test. The latter assumes knowledge of the standard deviation σ, whereas the former does not. In terms of the Key Inferential Function, this difference is plainly visible, whereas in traditional asymptotic studies one is usually left with the impression that it does not matter for large sample sizes. We showed that this is only true for the so-called contiguous alternatives.

21

Two-sample comparisons

Two-sample comparisons of normal populations, one sample undergoing treatment and the other serving as control, remain some of the most commonly encountered and challenging problems in statistics. It is difficult when the populations have different variances, the *heteroscedastic* case. The simpler *homoscedastic* case was solved long ago and will be reviewed in passing. Another difficulty is conceptual: does one want to compare the populations by estimating the difference of their means, the raw effect, or does one want to make inferences about a standardized effect, which is the effect size relative to the dispersion of the populations? As we shall show, the evidence in the Welch (1938) statistic for testing the hypothesis of no effect in the heteroscedastic case has two parameters, one a standardized effect and the second a sampling design factor, depending on how well the ratio of sample sizes agrees with the (unknown) ratio of population standard deviations.

The model entails $X_1, \ldots, X_{n_1}$ i.i.d. $N(\mu_1, \sigma_1^2)$ independent of $Y_1, \ldots, Y_{n_2}$ i.i.d. $N(\mu_2, \sigma_2^2)$, all parameters unknown. The objective is to test a null hypothesis $\mu_1 = \mu_2$ against a one-sided alternative, say $\mu_1 < \mu_2$, or the two-sided alternative. By definition the *effect* is $\theta = \mu_2 - \mu_1$ and the *standardized effect* is $\delta = \theta/\sigma$, where σ is a measure of scale, or magnitude, defined by (21.1). But first we consider an example, to help fix ideas.

21.1 Drop in systolic blood pressure

Summary statistics from the review by Mulrow *et al.* (2004) are shown in Table 21.1. The drop in systolic blood pressure following a weight-reducing diet for a group of patients was compared to that of a control group. The same article also includes reviews of studies that include hypertensive reducing drugs, and the interested reader

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

Table 21.1 Seven studies comparing drop in systolic blood pressure for treated patients undergoing a weight-loss regime (summarized by n_2, $\bar{y}$, s_2) with control patients not undergoing a weight-loss regime (summarized by n_1, $\bar{x}$, s_1). The estimated standardized effect in the kth study is $\hat{\delta}_k = \hat{\theta}_k/\hat{\sigma}_k = (\bar{y}_k - \bar{x}_k)/\hat{\sigma}_k$.

k	n_{1k}	$\bar{x}_k$	s_{1k}	n_{2k}	$\bar{y}_k$	s_{2k}	N_k	$\hat{\theta}_k$	$\hat{\delta}_k$
1	24	0.2	13.8	27	−4.8	13.8	51	−5.0	−0.18
2	18	7.4	8.1	20	13.3	8.1	38	5.9	0.36
3	64	4.0	15.7	66	11.0	17.1	130	7.0	0.21
4	9	−3.0	13.5	10	4.0	15.3	19	7.0	0.24
5	25	15.0	16.5	24	8.0	20.4	49	−7.0	−0.19
6	5	2.5	5.1	5	9.8	7.1	10	7.3	0.59
7	14	9.9	6.4	19	12.5	6.3	33	2.6	0.20

can find much more information on the selection criteria and methodology for these meta analyses by consulting the Cochran Review website at www.nicsl.com.au.

These data suggest homoscedasticity within each study, which will allow us to make comparisons of new techniques with traditional ones based on the assumption of equal variances. Our current objective is to find the evidence against each hypothesis $\theta_k = \mu_{2k} - \mu_{1k} = 0$ in the direction of $\theta_k > 0$ and to find a confidence interval for θ_k or δ_k. In Chapter 25 we will combine the evidence in all these studies and use it to find a confidence interval for a representative standardized effect.

First we need to consider what we mean by a standardized effect in one study. In the fifth study shown in Table 21.1, there is a negative effect −7.0, but the sample standard deviations of control and treatment groups are 16.5 and 20.4, so the effect is small relative to the spread within the control and treatment groups. By comparison, in the sixth study a positive effect of 7.3 is a little larger than the standard deviation in each group. Clearly the raw effects, in themselves, do not convey the discrepancy between control and treatment groups.

21.2 Defining the standardized effect

Writing $N = n_1 + n_2$ and $q = n_2/N$, the unbiased estimator of θ defined by $\hat{\theta} = \bar{Y}_{n_2} - \bar{X}_{n_1}$ has variance

$$\begin{aligned}\mathrm{Var}[\hat{\theta}] &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \\ &= \frac{1}{N}\left\{\frac{q\,\sigma_1^2 + (1-q)\,\sigma_2^2}{q(1-q)}\right\}. \end{aligned} \tag{21.1}$$

Letting σ^2 denote the quantity in the braces, we have $\mathrm{SE}[\hat{\theta}] = \sigma/\sqrt{N}$, where σ is our chosen measure of scale; it is free of N and depends only on the known relative sample sizes and the unknown population variances.

When $\sigma_1 = \sigma_2$, the standard error reduces to $\sigma/\sqrt{N} = \sigma_1/\sqrt{N\{q(1-q)\}}$, a quantity minimized by taking $q = 0.5$, and then $\sigma = 2\sigma_1$. More generally, Equation (21.1) is minimized for fixed N by choosing the second sample proportion to be $q = \sigma_1/(\sigma_1 + \sigma_2)$. This will help if one has a rough idea of the relative sizes of σ_1, σ_2 prior to sampling, and wants to estimate the effect θ.

We define the *standardized effect* by $\delta = \theta/\sigma$. In this chapter we find confidence intervals for δ and evidence for $\delta > 0$. When $\sigma_1 = \sigma_2$, the standardized effect reduces to $\delta = \{q(1-q)\}^{1/2}\{(\mu_2 - \mu_1)/\sigma_1\}$. The second factor is often referred to as Cohen's-*d* (Cohen 1988) in the psychological literature, and is called the effect size in Hedges and Olkin (1985). The first factor $\{q(1-q)\}^{1/2}$ is known, and reflects the impact of unbalanced sampling. We will not assume equal variances further. It is helpful to rewrite δ in terms of $q = n_2/N$, $\varrho = \sigma_2^2/\sigma_1^2$ and $\delta_1 = \theta/\sigma_1$:

$$\delta = \delta_1 \left\{ \frac{1}{1-q} + \frac{\varrho}{q} \right\}^{-1/2} . \tag{21.2}$$

The ratio $\delta_1 = \theta/\sigma_1$ is the standardized effect relative to the scale σ_1 of the first (control) sample; it was originally proposed by Glass (1976), who argued that the control sample should be the basis for standardization. For balanced sampling, that is, $q = 0.5$, the standardized effect $\delta = \delta_1(2 + 2\varrho)^{-1/2}$.

21.3 Evidence in the Welch statistic

It is clear that testing the hypotheses $\theta = 0$ versus $\theta > 0$ is equivalent to testing $\delta = 0$, versus $\delta > 0$, so in either case the evidence for the alternative will be the same. To find it, we begin by variance stabilizing the Welch statistic, which is widely used for testing these hypotheses in the context of comparing two normal populations, with all four parameters unknown.

21.3.1 The Welch statistic

Welch (1938) proposed a test statistic for $\delta = 0$ versus $\delta > 0$ which is defined by $t_{\text{Welch}} = \sqrt{N}(\bar{Y}_{n_2} - \bar{X}_{n_1})/\hat{\sigma}$, where $\hat{\sigma}$ is the estimate of σ obtained by substituting the sample variances s_1^2, s_2^2 for the respective population variances σ_1^2, σ_2^2. Welch (1938, 1947) and Aspin (1948) showed that the distribution of t_{Welch} under the null $\delta = 0$ is approximately the Student t-distribution with ν degrees of freedom. It has df $\nu = (A + B)^2/\{A^2/(n_1 - 1) + B^2/(n_2 - 1)\}$, where $A = \sigma_1^2/n_1$, $B = \sigma_2^2/n_2$. In implementation, ν is estimated by $\hat{\nu}$ obtained by substitution of s_1^2, s_2^2 for the unknown population parameters. Further, an approximate $100(1-\alpha)$ % confidence interval for θ is given by

$$[L, U] = \left[\hat{\theta} - t_{\hat{\nu},1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{N}} , \; \hat{\theta} + t_{\hat{\nu},1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{N}} \right]. \tag{21.3}$$

The statistic t_{Welch} can be written as the ratio of independent variables, $Z + \delta$ to W, where Z is standard normal and W is the square root of a mixture of independent $\chi^2_{n_1-1}$, $\chi^2_{n_2-1}$ variables. The exact distribution of t_{Welch} for all parameter values is derived in Nel *et al.* (1990), and is a generalization of the noncentral F-distribution. They use it to show that Welch's approximate t-distribution methodology is quite accurate for obtaining critical points. It does not appear to be useful in obtaining confidence intervals for δ because of its complicated form and dependence on the unknown population variances.

21.3.2 Variance stabilizing the Welch *t*-statistic

Let $w = 1/\text{Var}[\hat{\theta}]$ be the inverse of the variance of the effect estimator $\hat{\theta} = \bar{Y}_{n_2} - \bar{X}_{n_1}$. Further define the constants $C = \sigma_1^4/\{n_1^2(n_1-1)\} + \sigma_2^4/\{n_2^2(n_2-1)\}$ and $D = \sigma_1^6/\{n_1^3(n_1-1)^2\} + \sigma_2^6/\{n_2^3(n_2-1)^2\}$. Then it is shown in Kulinskaya and Staudte (2007) that

$$\text{E}[t_{\text{Welch}}] \approx \sqrt{N}\,\delta \left\{1 + \frac{3w^2C}{4} + K\left[\frac{105w^4C^2}{32} - \frac{5w^3D}{2}\right]\right\} \tag{21.4}$$

$$\text{Var}[t_{\text{Welch}}] \approx 1 + 2w^2C + N\,\delta^2 \left\{\frac{w^2C}{2} + K\left[\frac{39w^4C^2}{8} - 3w^3D\right]\right\}. \tag{21.5}$$

When $K = 0$ the terms in curly brackets are accurate to order $O(N^{-1})$, and when $K = 1$ to order $O(N^{-2})$. The choice of K will be made later to improve the range of parameters for which 95 % confidence intervals for δ are obtained.

It follows from these approximations that $\text{Var}[t_{\text{Welch}}] = a_1 + a_2\text{E}^2[t_{\text{Welch}}]$, where $a_1 = 1 + 2w^2C$ and a_2 is the ratio of the quantity in curly brackets in (21.5) to the square of the quantity in curly brackets in (21.4). It then follows by the same derivation used in Section 20.2 that a variance stabilizing transformation of the Welch statistic is given by:

$$T(t_{\text{Welch}}) = \frac{1}{a_2^{1/2}}\,\sinh^{-1}\left\{\left(\frac{a_2}{a_1}\right)^{1/2} t_{\text{Welch}}\right\}. \tag{21.6}$$

The theory suggests this transformed statistic $T = T(t_{\text{Welch}})$ should be approximately normally distributed with variance 1, but simulations are needed to verify these claims (see below). Letting n_1, n_2 tend to infinity with proportion $q = n_2/(n_1+n_2) = n_2/N$ fixed, it follows that $a_1 \to 1$ and $Na_2 \to \xi/2$, where

$$\xi = \lim \frac{N}{\nu} = \frac{\sigma_1^4/(1-q)^3 + \sigma_2^4/q^3}{\{\sigma_1^2/(1-q) + \sigma_2^2/q\}^2} = \frac{(1-q)^{-3} + \varrho^2q^{-3}}{\{(1-q)^{-1} + \varrho q^{-1}\}^2}. \tag{21.7}$$

Thus the variance stabilized Welch statistic in (21.6) will for large n_1, n_2 have expected value $\mathrm{E}[T] \doteq \sqrt{N}\mathcal{K}(\delta)$, where the Key is

$$\mathcal{K}(\delta) = \sqrt{\frac{2}{\xi}}\,\sinh^{-1}\left(\frac{\sqrt{\xi}\,\delta}{\sqrt{2}}\right). \tag{21.8}$$

Here ξ is a parameter depending only on the unknown ratio of variances ϱ and the proportion q of observations allocated to the second sample. For any $\delta > 0$ the expected evidence τ is decreasing in ξ, so to maximize τ we want ξ to be as small as possible. Using elementary calculus, one can show $\xi \geq 1$ with equality along the curve $\varrho = \{q/(1-q)\}^2$, or $q = \sqrt{\varrho}/(1+\sqrt{\varrho}) = \sigma_2/(\sigma_1+\sigma_2)$. This is the same 'choice' of q which minimizes the standard error of $\hat{\theta}$ (see the discussion after Equation (21.1)). Because the ratio of population variances is usually unknown, balanced sampling $q = 0.5$ is recommended, for then the constant ξ is bounded and in fact $1 \leq \xi \leq 2$.

When $\xi = 1$ the asymptotic mean (21.8) reduces to $\tau = \sqrt{N}\,\mathcal{K}(\delta)$, where $\mathcal{K}(\delta)$ is given by (20.5); that is, τ reduces to the same expression for the mean evidence encountered in the one-sample Student t-statistic. This is in accord with the homogeneous case $\varrho = 1$ and balanced sampling, for then the Welch statistic and the two-sample pooled t-test are approximately equal. Even in the nonhomogeneous case, the Welch test statistic with $q = \sigma_2/(\sigma_1+\sigma_2)$ has $\nu \approx N$. Further, the scale parameter defined in (21.1) simplifies to $\sigma = \sigma_1 + \sigma_2$. Thus in this case the scale parameter depends on only one unknown, the sum of the standard deviations, so it is not surprising that the evidence in t_{Welch} is then asymptotically equivalent to that in a one-sample t-statistic. In the next section we describe the small-sample behavior of the variance stabilized t_{Welch}.

21.3.3 Choosing the sample size to obtain evidence

For any fixed $0 < q < 1$ and $\xi = \xi(\varrho)$ defined by (21.7) one can show that $\xi(\varrho) \leq \xi_{\max} = \max_{0<\varrho<\infty} \xi(\varrho) = \max\{q^{-1}, (1-q)^{-1}\}$. Further the expected evidence (21.8) in t_{Welch} has, for any fixed $\delta_1 > 0$, its smallest magnitude when $\xi = \xi_{\max}$, because $|\tau|$ is decreasing in ξ. Thus to guarantee expected evidence of at least $|\tau_1|$ for all ϱ when $\delta = \delta_1$ we need to choose N_1 so that

$$N_1 = \frac{\tau_1^2\,\xi_{\max}}{2} \bigg/ \left[\sinh^{-1}\left(\delta_1\sqrt{\frac{\xi_{\max}}{2}}\right)\right]^2. \tag{21.9}$$

For balanced sampling $N = 2n_1$ this reduces to $N_1 = \left\{\tau_1/\sinh^{-1}(\delta_1)\right\}^2$.

21.4 Confidence intervals for δ

21.4.1 Converting the evidence to confidence intervals

If it is true that $T = T(t_{\text{Welch}})$ defined in (21.6) satisfies $T \sim N(\tau, 1)$ at least approximately, with τ given by (21.8), then a nominal $100(1-\alpha)$ % confidence interval for τ

is $[T - z_{1-\alpha/2}, T + z_{1-\alpha/2}]$. Now $\tau = \sqrt{2N/\xi}\,\sinh^{-1}(\sqrt{\xi}\,\delta/\sqrt{2})$, so this confidence interval for τ can be modified to isolate δ between two limits, namely

$$\left[\sqrt{\frac{2}{\hat{\xi}}}\,\sinh\left\{\sqrt{\frac{\hat{\xi}}{2N}}\,(T - z_{1-\alpha/2})\right\},\ \sqrt{\frac{2}{\hat{\xi}}}\,\sinh\left\{\sqrt{\frac{\hat{\xi}}{2N}}\,(T + z_{1-\alpha/2})\right\}\right], \quad (21.10)$$

where $\hat{\xi}$ is an estimate of ξ. This is a nominal $100(1-\alpha)$ % confidence interval for δ, but needs to be checked by simulations. In the following we substitute sample variances for population variances in (21.7) to obtain $\hat{\xi}$.

21.4.2 Simulation studies

Kulinskaya and Staudte (2007) report on a series of simulation studies for different choices of $\varrho = \sigma_2^2/\sigma_1^2$ and sample proportions $q = n_2/(n_1 + n_2)$ with total sample size $N = n_1 + n_2$ ranging from 10 to 100, and raw effect $\theta = \mu_2 - \mu_1$ ranging from 0 to 5. For each configuration (q, r, N, θ), 100 000 values of t_{Welch} were generated, and $T = T(t_{\text{Welch}})$ computed using transformation (21.6). After experimentation, the constant $K(q) = \min\{1, 1/3 + 4\,|q - 1/2|\}$ required in (21.4) and (21.5) was found to improve the accuracy of 95 % confidence intervals for δ. The empirical means of $T/\sqrt{N}$ are close to the asymptotic value $\tau/\sqrt{N}$ and the empirical standard deviation of T is also close to 1 for a wide range of parameter values. For complete details we refer the interested reader to Kulinskaya and Staudte (2007).

Table 21.2 gives a summary of the coverage probability results of nominal 95 % confidence intervals over the range $0 \le \delta \le 1$. In the context of equal variances $\sigma_1^2 = \sigma_2^2$, Cohen (1988) has described effect size values $d = (\mu_2 - \mu_1)/\sigma_1$ equal to 0.2, 0.5 and 0.8 as 'small', 'medium' and 'large', respectively. These correspond to our $\delta = 0.1, 0.25$ and 0.4. Hedges and Olkin (1985, p. 87) similarly note that effect size

Table 21.2 Columns 2–4 list the minimum and maximum empirical coverage percentages of the nominal 95 % confidence intervals (21.10) for δ, when $0 \le \delta \le 1$, based on 100 000 simulations. All numbers are rounded to the nearest 0.1 %.

(n_1, n_2)	$\varrho = 1$	$\varrho = 2$	$\varrho = 4$
(5, 5)	(95.2, 95.5)	(95.0, 95.2)	(94.5, 94.7)
(10, 10)	(95.1, 95.2)	(95.1, 95.2)	(94.6, 95.0)
(20, 20)	(95.0, 95.2)	(95.0, 95.2)	(94.7, 95.2)
(25, 25)	(95.1, 95.2)	(94.9, 95.2)	(94.8, 95.1)
(50, 50)	(94.9, 95.2)	(95.0, 95.1)	(94.9, 95.1)
(100, 100)	(94.9, 95.2)	(94.9, 95.1)	(94.8, 95.2)
(5, 10)	(94.8, 95.6)	(95.1, 95.6)	(95.2, 95.8)
(10, 20)	(95.0, 95.2)	(95.0, 95.3)	(95.1, 95.5)
(20, 40)	(95.0, 95.1)	(95.0, 95.2)	(95.0, 95.3)
(30, 60)	(94.9, 95.0)	(94.9, 95.2)	(95.0, 95.1)
(60, 120)	(94.9, 95.0)	(95.0, 95.1)	(95.0, 95.1)

Table 21.3 Values of the t_{Welch} statistic, the variance stabilized statistic $T = T(t_{\text{Welch}})$ given by (21.6) and 95 % confidence intervals $[L_i, U_i]$ for δ_i based on (21.10) and the data given in Table 21.1.

i	N_i	$\hat{\theta}_i$	$\hat{\sigma}_i/\sqrt{N_i}$	$t_{\text{Welch,i}}$	T_i	$[L_i, U_i]$
1	51	−5.0	3.87	−1.29	−1.26	[−0.46, +0.10]
2	38	5.9	2.63	2.24	2.16	[+0.03, +0.70]
3	130	7.0	2.88	2.43	2.40	[+0.04, +0.39]
4	19	7.0	6.61	1.06	1.00	[−0.22, +0.71]
5	49	−7.0	5.31	−1.31	−1.29	[−0.47, +0.10]
6	10	7.3	3.91	1.87	1.61	[−0.11, +1.37]
7	33	2.6	2.24	1.16	1.12	[−0.15, +0.55]

values from quantitative research syntheses usually fall within their considered range of 0 to 1.5, which corresponds to our $0 \leq \delta \leq 0.75$. It is also found in Kulinskaya and Staudte (2007) that the lengths of these confidence intervals for δ, which do not assume equal variances, are only slightly longer than the best available when the assumption is made.

21.4.3 Drop in systolic blood pressure (continued)

Table 21.3 shows the results of applying the procedures proposed in Sections 21.3 and 21.4.1 to each of the individual studies in Table 21.1. The evidence for $\delta > 0$ is negligible in all studies except 2, 3 and 6 where it is weak.

Only the second and third studies are of level 0.05 significance in testing $\delta_i = 0$ against $\delta_i \neq 0$, because the 95 % confidence intervals for δ_2, δ_3 shown in column 7 do not contain 0. The same two studies would yield 0.05 significance for testing $\mu_{2i} = \mu_{1i}$ against $\mu_{2i} \neq \mu_{1i}$ using the traditional Welch t intervals for the differences in means. So the question should be asked: what have we gained by considering standardized effects rather than the raw differences in treated and control responses?

What we have gained is the ability to make comparisons among studies, because the standardized effects are free of the variability in the populations under consideration in all the studies. Moreover, we can put all seven results together to combine the evidence for various alternatives and to estimate an overall effect (see Chapter 25).

21.5 Summary

The evidence in the two-sample Welch statistic is found to have a form similar to that in the one-sample t-statistic, again growing with a suitably defined standardized effect through the $\sinh^{-1}$ transformation, but now also dependent on a parameter ξ which is a known function of the relative sample sizes and the unknown ratio of variances. Knowing this enables one to choose the sample size to obtain a desired expected evidence for a given δ.

For any fixed $\delta > 0$ the expected evidence in t_{Welch} has a maximum near that of a one-sample t-test with N degrees of freedom when $q = \sigma_2/(\sigma_1 + \sigma_2)$ because then $\xi = 1$. Larger values of ξ will only diminish the expected evidence in the test statistic. Because the ratio of variances is usually unknown, it is recommended that sampling be balanced, for then at least $1 \leq \xi \leq 2$.

Confidence intervals for the unknown standardized effect can be obtained for a wide range of parameter values and even small sample sizes, provided sampling is balanced.

22

Evidence in the chi-squared statistic

22.1 The noncentral chi-squared distribution

A comprehensive collection of results on the noncentral chi-squared distribution is found in Johnson *et al.* (1995, Chapter 29); most material in this section is excerpted from their work but stated in our notation.

Definition 22.1 *Given ν independent standard normal variables $Z_1, \ldots, Z_\nu$ and ν constants $\mu_1, \ldots, \mu_\nu$, the distribution of $S = \sum_{i=1}^{\nu}(Z_i + \mu_i)^2$ has the noncentral chi-squared distribution, and depends only on two parameters, ν and $\lambda = \sum_{i=1}^{\nu} \mu_i^2$. It is denoted $S \sim \chi_\nu^2(\lambda)$. When $\lambda = 0$ this distribution is the standard (central) chi-squared distribution, denoted χ_ν^2.*

In many applications the null $\lambda = 0$ is rejected at level α in favor of the alternative $\lambda > 0$ when $S \geq \chi^2_{\nu,1-\alpha}$, the $1 - \alpha$ quantile of the χ^2_ν distribution. Denote the null median by $m_\nu = \chi^2_{\nu,0.5}$. The difference $\nu - m_\nu$ between the null mean and median is monotonically increasing with ν from a minimum of 0.545 at $\nu = 1$ to a least upper bound of 2/3. The mean and variance of $S \sim \chi_\nu^2(\lambda)$ are given by $\mathrm{E}[S] = \nu + \lambda$ and $\mathrm{Var}[S] = 2\nu + 4\lambda$.

Example. Between group sum of squares (for known variance)

For each group $k = 1, \ldots, K$ let $\mathbf{X}_k' = [X_{k1}, X_{k2}, \ldots, X_{k,n_k}]$ denote a sample of n_k observations, each with distribution $N(\mu_k, 1)$. Also assume the elements of

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

$\mathbf{X}' = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$ are independent. Further introduce the total sample size $N = \sum_k n_k$, the sample proportions $q_k = n_k/N$, the kth sample mean $\bar{X}_k$, the overall sample mean $\bar{X} = \sum_k q_k \bar{X}_k$, its expectation $\mu = \sum_k q_k \mu_k$ and the parameter $\lambda = N \sum_k q_k(\mu_k - \mu)^2$. Then the between group sum of squares $Y = N \sum_k q_k(\bar{X}_k - \bar{X})^2 \sim \chi^2_\nu(\lambda)$, where $\nu = K - 1$. The ratio $\theta = \lambda/N = \sum_k q_k(\mu_k - \mu)^2$ depends only on the *relative* sample sizes q_k, and measures the variability of the group means μ_k using a weighted sum of squared deviations from the weighted mean μ, with weights q_k.

To see why $Y \sim \chi^2_\nu(\lambda)$, for every positive integer n denote by $\mathbf{1}_n$ the n-vector of 1's, I_n the $n \times n$ identity matrix and J_n the $n \times n$ matrix of 1's. Let M_N be the block diagonal matrix with kth diagonal submatrix J_{n_k}/n_k. Then for $C = M_N - J_N/N$ the between group sum of squares can be written as the quadratic form $Y = \mathbf{X}'C\mathbf{X}$ where C is symmetric and idempotent. Hence by Rao (1973, Section 3.b.4) $Y \sim \chi^2_\nu(\lambda)$, where $\nu = K - 1$ is the rank of C (equal to the trace of C) and the noncentrality parameter $\lambda = \mathrm{E}[\mathbf{X}]' \, C \, \mathrm{E}[\mathbf{X}] = N \sum_k q_k(\mu_k - \mu)^2$.

22.2 A *vst* for the noncentral chi-squared statistic

The asymmetry in the chi-squared distribution means that one must work harder to stabilize its variance, and that the Key Inferential Function will depend on the sample size. Nevertheless, the methodology can be carried out with useful consequences for inference.

22.2.1 Deriving the *vst*

Let S be any test statistic with $S \sim \chi^2_\nu(\lambda)$. We want a *vst* $T_\nu = h_\nu(S)$ such that T_ν satisfies the properties E_1 to E_4 of Section 16.1.1; that is, ideally it should be monotonically increasing in S, have variance 1, be normally distributed for all λ and have expectation monotonically increasing from 0 at the null with λ.

Now $\mathrm{E}[S] = \nu + \lambda$ and $\mathrm{Var}[S] = 2\nu + 4\lambda = g(\mathrm{E}[S])$, where $g(t) = 4t - 2\nu$, so by the method of Section 17.2, $h(x) = \int^x \{g(t)\}^{-1/2} \mathrm{d}t = \sqrt{x - \nu/2} + c_\nu$, where c_ν is any constant, should stabilize the variance of $h(S)$ near 1. Further, the choice $c_\nu = -\sqrt{\nu/2}$, yields a first-order expected evidence $\mathrm{E}[h(S)] \doteq \sqrt{\lambda + \nu/2} - \sqrt{\nu/2}$, which is monotonically increasing from 0 with λ.

However, the above promising heuristic argument is flawed, because $\mathrm{E}[S] \geq \nu$ and therefore the relationship $\mathrm{Var}[S] = g(\mathrm{E}[S])$ underlying the derivation is defined only for $t \geq \nu$, and hence $s \geq \nu$. The definition of the *vst* needs to be extended to small values of S in a smooth way so that the properties E_1 to E_4 are satisfied. This can be done if one centers the *vst* on the null median m_ν instead of the null mean ν, and then defines the *vst* for small S in terms of large S by a symmetrization about m_ν as follows.

Definition 22.2 *Let F_ν be the cdf of the central chi-squared distribution with ν degrees of freedom and let m_ν be the median of this distribution: $F_\nu(m_\nu) = 0.5$. For the model $S \sim \chi^2_\nu(\lambda)$, $\lambda \geq 0$ and hypotheses $\lambda = 0$ against $\lambda > 0$, the evidence in S is* *defined by*

$$T_\nu = h_\nu(S) = \begin{cases} +\sqrt{S \ - m_\nu/2} \ - \sqrt{m_\nu/2}, & \textit{for } S \geq m_\nu; \\ -\sqrt{S^* - m_\nu/2} \ + \sqrt{m_\nu/2}, & \textit{for } S < m_\nu. \end{cases} \qquad (22.1)$$

where $S^* = F_\nu^{-1}(1 - F_\nu(S))$. *Positive values of* T_ν *are interpreted as evidence for the hypothesis* $\lambda > 0$. *Negative values of* T_ν *will be interpreted, after multiplication by minus one, as positive evidence for the null hypothesis* $\lambda = 0$.

Some graphs of T_ν are plotted in Figure 22.1 which reveal T_ν to be a smooth, monotonically increasing function of S that is 0 at the null median. Further, as simulations reported in Section 22.3 demonstrate, the *vst* defined by (22.1) has variance well-stabilized near 1 and is approximately normal for a wide range of parameter values. Thus this *vst* will be seen to satisfy E_1 to E_3 of Section 16.1.1.

22.2.2 The Key Inferential Function

It turns out that the first-order mean $h_\nu(\mathrm{E}[S])$ is too rough an approximation for our applications, and the bias correction term in (17.1) is also needed. In terms of the parameters $\theta = \lambda/N$, $\nu = K - 1$, the mean and variance of S are $\mathrm{E}[S] = N\{\theta + \nu/N\}$ and $\mathrm{Var}[S] = N\{4\theta + 2\nu/N\}$. Hence Equation (17.1) leads to the following key.

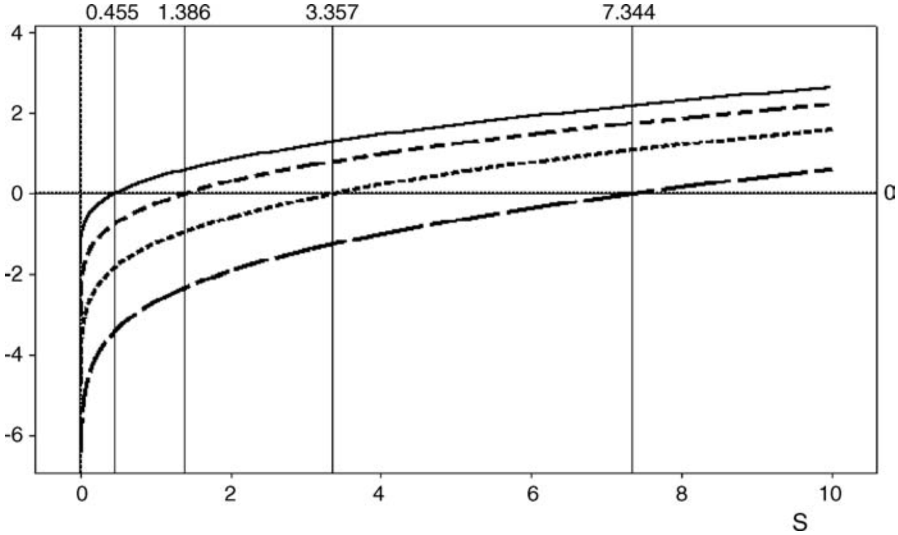


Figure 22.1 The graph of $T = T(S)$ defined by (22.1) is shown for $\nu = 1$ as a solid line, $\nu = 2$ as a dashed line, $\nu = 4$ as as dotted line and $\nu = 8$ as a long-dashed line. The vertical reference lines show the respective central chi-squared medians.

Definition 22.3 *The Key Inferential Function for the noncentral chi-squared model when testing $\theta = 0$ against $\theta > 0$ is defined by*

$$\mathcal{K}_{N,\nu}(\theta) = \mathcal{L}_{N,\nu}(\theta) - \frac{\left(\theta + \frac{\nu}{2N}\right)}{2N\mathcal{L}^3_{N,\nu}(\theta)} - \left\{\frac{m_\nu}{2N}\right\}^{1/2}, \tag{22.2}$$

where

$$\mathcal{L}_{N,\nu}(\theta) = \left\{\theta + \frac{2\nu - m_\nu}{2N}\right\}^{1/2}. \tag{22.3}$$

A simpler, but less accurate, Key than (22.2) is obtained by replacing ν in its second term by $2\nu - m_\nu$ to obtain

$$\mathcal{K}^*_{N,\nu}(\theta) = \mathcal{L}_{N,\nu}(\theta) - \frac{1}{2N\mathcal{L}_{N,\nu}(\theta)} - \left\{\frac{m_\nu}{2N}\right\}^{1/2}. \tag{22.4}$$

Remarks

1. The Key Inferential Function defined by (22.2) for parameters $N \geq \nu \geq 1$ and $\theta = \lambda/N$ of practical interest, gives the expected evidence in the noncentral chi-squared distribution, in the sense that (22.1) satisfies
$$\mathrm{E}[T_\nu] \doteq \sqrt{N}\,\mathcal{K}_{N,\nu}(\theta). \tag{22.5}$$
2. For $\nu = K - 1$ fixed and $\lambda, N \to \infty$ with $\theta = \lambda/N$ fixed, $\mathcal{K}_{N,\nu}(\theta) \to \theta^{1/2}$. This remains true even if the number of groups $K = \nu + 1$ grows with N at any rate less than N; that is, $K = o(N)$.
3. There are some applications where one wants to test the hypotheses $\lambda < \lambda_0$ versus $\lambda > \lambda_0$, where the boundary λ_0 is positive. A vertical adjustment to the evidence (22.1) allows one to do this (see Section 22.5).

22.3 Simulation studies

In order to assess the properties of T_ν defined by (22.1), 400 000 samples were generated for various degrees of freedom $\nu \geq 2$ and values of $\theta = \lambda/N$ ranging from 0 to 3 in steps of 0.1. For example, with $K = 3$ groups, $\nu = K - 1 = 2$, and for a total number of observations $N = 6,\ 12$ and 24, the results are displayed in Figure 22.2. The simulations for $\nu = 2$, but with $N = 50,\ 100$ and 200, are shown in Figure 22.3. Details now follow.

22.3.1 Bias in the evidence function

The bias $E[T_\nu] - \sqrt{N}\,\mathcal{K}_{N,\nu}(\theta) = \sqrt{N}\,\{E[T_\nu]/\sqrt{N} - \mathcal{K}_{N,\nu}(\theta)\}$ of T_ν in estimating $\sqrt{N}$ times the Key inferential Function in (22.2) is shown as a function of $0 \leq \theta \leq 3$ for $N = 6,\ 12$ and 24 in the top left-hand plot of Figure 22.2. The maximum absolute bias (absolute difference between the two sides of (22.5)) is less than 0.04,

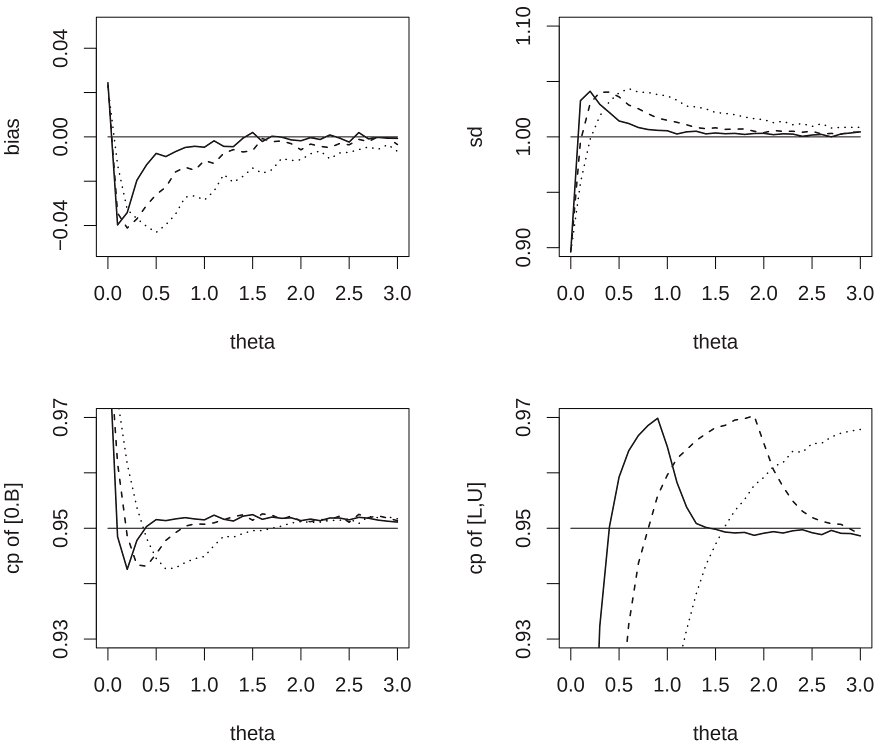


Figure 22.2 In the first row of plots are shown the empirical biases and standard deviations of T_2 for $N = 6$ (dotted line), $N = 12$ (dashed line) and $N = 24$ (solid line). The second row of plots gives the empirical coverage probabilities of nominal 95 % upper confidence bounds and 95 % conditional confidence intervals.

so the bias squared is much smaller than the variance of T_2, which is close to 1, as illustrated in the top right-hand plot of the same figure. The distributions of T_2 are very close to normality (not shown). These properties suggest that confidence bounds and confidence intervals for θ can be found for $\theta = \lambda/N$ and hence λ.

22.3.2 Upper confidence bounds; confidence intervals

To the extent that T_ν satisfies properties E_1 to E_4 of a measure of evidence, one can expect $T_\nu + z_{0.95}$ to define a nominal 95 % upper confidence bound for $\sqrt{N}\,\mathcal{K}_{N,\nu}(\theta)$, and hence $\left[0,\ \mathcal{K}_{N,\nu}^{-1}\left(\{T_\nu + z_{0.95}\}/\sqrt{N}\,\right)\right]$ defines a nominal 95 % upper confidence bound for θ. However, an explicit formula for $\mathcal{K}_{N,\nu}^{-1}$ is not readily obtained, so we based our confidence bound on $\mathcal{K}_{N,\nu}^{*}$ of (22.4).

Definition 22.4 *For fixed* $\alpha < 0.5$ *define* $U_{\nu,\alpha} = \left(T_\nu + z_{1-\alpha} + \sqrt{m_\nu/2}\,\right)^2$*. Then, after inversion of* $\mathcal{K}_{N,\nu}^{*}$*, which requires the solving of a quadratic equation, one obtains a*

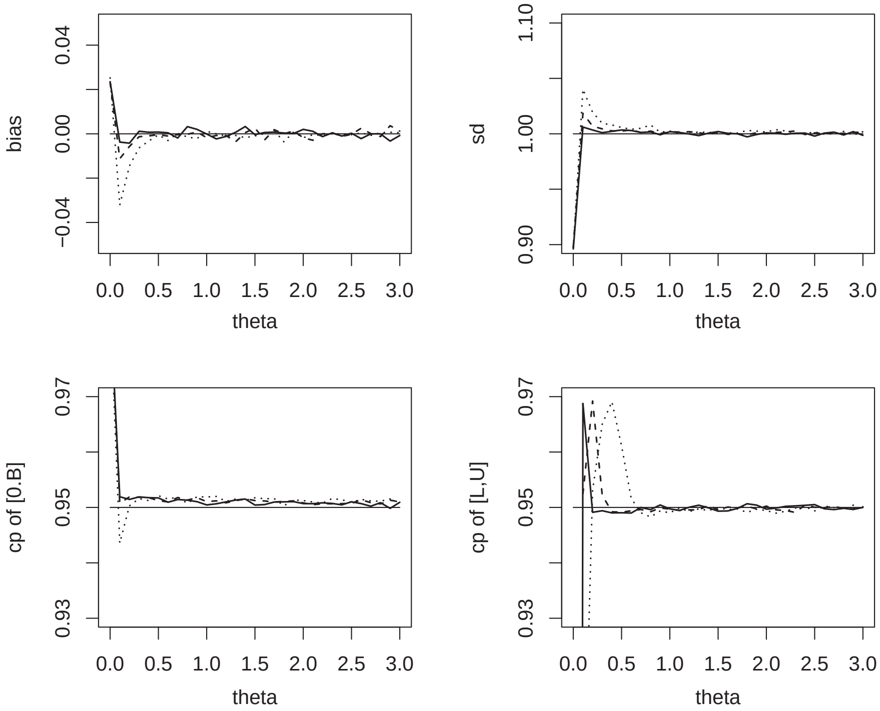


Figure 22.3 A repeat of Figure 22.2, but now for $N = 50$ (dotted line), $N = 100$ (dashed line) and $N = 200$ (solid line).

nominal upper 100$(1 - \alpha)$ *% confidence bound for* θ *of the form* $[0, B]$ *defined by*

$$B = \frac{U_{\nu,\alpha}}{4N}\left(1 + \left\{1 + \frac{2}{U_{\nu,\alpha}}\right\}^{1/2}\right)^2 - \frac{\nu}{2N}. \tag{22.6}$$

The reader who carries out the derivation will find that instead of the constant $-\nu/(2N)$ appearing on the right-hand side of (22.6), one has $(m_\nu - 2\nu)/(2N)$. The change to $-\nu/(2N)$ brings the coverage probability closer to the nominal value when $1 - \alpha = 0.95$, although sometimes it drops slightly below 95 %. To ensure at least 95 % confidence for all θ we recommend replacing $-\nu/(2N)$ with $-m_\nu/(2N)$. For examples of the empirical coverage probabilities of nominal 95 % upper confidence bounds when $\nu = 2$, see Figures 22.2 and 22.3.

Reliable confidence intervals for θ are more difficult to obtain because the bias and variance of T_ν are not close to 0 and 1, respectively, when θ is small. However, if one only tries to form an interval when T_ν exceeds $z_{1-\alpha/2}$, that is, when T_ν is large enough to be significant at level $\alpha/2$ for alternative $\theta > 0$, then one can expect some success. Therefore we define conditional confidence intervals as follows.

Definition 22.5 *Assume that* $T_\nu > z_{1-\alpha/2}$. *Then subject to this condition, define* $L_{\nu,\alpha/2} = \left(T_\nu - z_{1-\alpha/2} + \sqrt{m_\nu/2}\ \right)^2$ *and* $U_{\nu,\alpha/2} = \left(T_\nu + z_{1-\alpha/2} + \sqrt{m_\nu/2}\ \right)^2$. *A nominal 100*$(1-\alpha)$ *% (conditional) confidence interval* $[L, U]$ *for* θ *has endpoints*

$$L = \frac{L_{\nu,\alpha/2}}{4N}\left(1 + \left\{1 + \frac{2}{L_{\nu,\alpha/2}}\right\}^{1/2}\right)^2 - \frac{m_\nu}{2N}, \tag{22.7}$$

$$U = \frac{U_{\nu,\alpha/2}}{4N}\left(1 + \left\{1 + \frac{2}{U_{\nu,\alpha/2}}\right\}^{1/2}\right)^2 - \frac{m_\nu}{2N}. \tag{22.8}$$

Again, the choice of additive constant $-m_\nu/2N$ in these equations yields empirical coverage closer to the 95 % value. Some empirical coverage probabilities of these intervals are displayed in Figures 22.2 and 22.3.

More examples are shown in Figures 22.4 and 22.5. In Figures 22.4 there were $K = 5$ groups, and hence $\nu = 4$, with total number of observations $N = 10$, 20

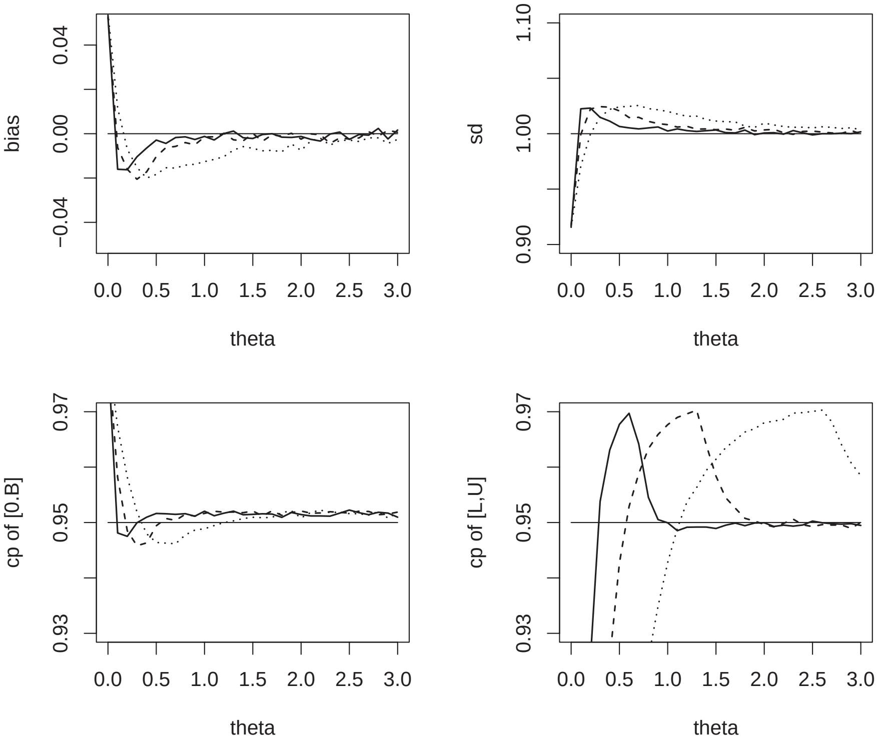


Figure 22.4 In the first row of plots are shown the empirical biases and standard deviations of T_4 for $N = 10$ (dotted line), $N = 20$ (dashed line) and $N = 40$ (solid line). The second row of plots gives the empirical coverage probabilities of nominal 95 % upper confidence bounds and 95 % conditional confidence intervals.

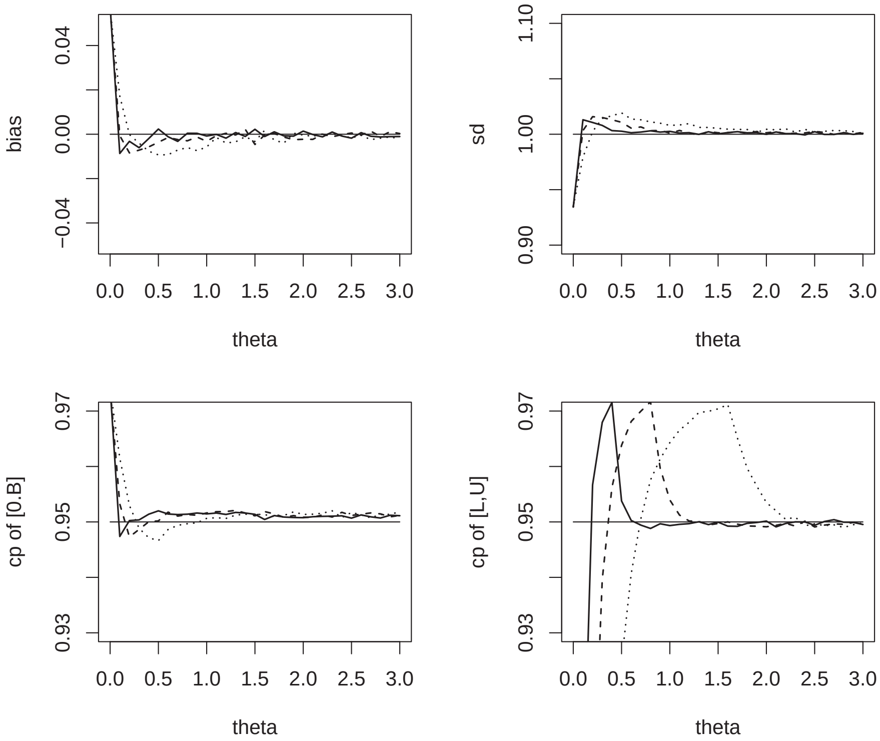


Figure 22.5 The corresponding results for $\nu = 9$, and $N = 20$ (dotted line), $N = 40$ (dashed line) and $N = 80$ (solid line).

and 40. Figure 22.5 displays the results for $K = 10$ groups, so $\nu = 9$ and here $N = 20,\ 40$ and 80. Note that the conditional confidence intervals are not very reliable in terms of coverage unless θ is moderately large.

22.4 Choosing the sample size

22.4.1 Sample sizes for obtaining an expected evidence

Earlier in Section 17.2.2 some useful properties of the Key Inferential Function were listed. In particular, property K_1 states that if one wants expected evidence τ_1 for alternative θ_1, one needs to solve for the least integer N satisfying $N \geq \{\tau_1/\mathcal{K}_{N,\nu}(\theta_1)\}^2$. This goal is more readily accomplished by solving for N in $\sqrt{N}\,\mathcal{K}^*_{N,\nu}(\theta_1) = \tau_1$, where $\mathcal{K}^*_{N,\nu} \doteq \mathcal{K}_{N,\nu}$ is found in (22.4). The equation of interest can now be rewritten in terms of $\lambda_1 = N\theta_1$, $a = a_\nu = \nu - m_\nu/2$ and $b = b_\nu(\tau) = \tau_1 + \sqrt{m_\nu/2}$ as

$$\sqrt{\lambda_1 + a} - \frac{1}{2\sqrt{\lambda_1 + a}} = b, \tag{22.9}$$

which leads to a quadratic equation with positive solution

$$\lambda_1 \;=\; \lambda_1(\nu, \tau_1) \;=\; \frac{1 + b^2 + b\sqrt{2 + b^2}}{2} - a. \qquad (22.10)$$

It follows that the minimum sample size required to obtain expected evidence τ_1 for alternative θ_1 is the least integer N_1 greater than or equal to $\lambda_1(\nu, \tau_1)/\theta_1$. Dropping the subscripts, plots of the function $\lambda(\nu, \tau)$ against τ defined by (22.10) are shown in Figure 22.6, for $\nu = 1, \ldots, 9$. These plots make it easy to quickly determine the sample size required to obtain expected evidence τ_1 for alternative θ_1.

For example, suppose we want moderate expected evidence of $\tau_1 = 3.3$ for alternative $\theta_1 = 0.5$. By following the vertical dashed line in Figure 22.6 up to the graph for $\nu = 4$, and then the horizontal dashed line over to the y-axis, one finds $\lambda(4, 3.3) = 19.8$. This leads to the minimum sample size $N_1 = 19.8/0.5 \approx 40$. For the same expected evidence against the smaller alternative $\theta_1 = 0.2$, one would need a sample size $N_1 \approx 100$.

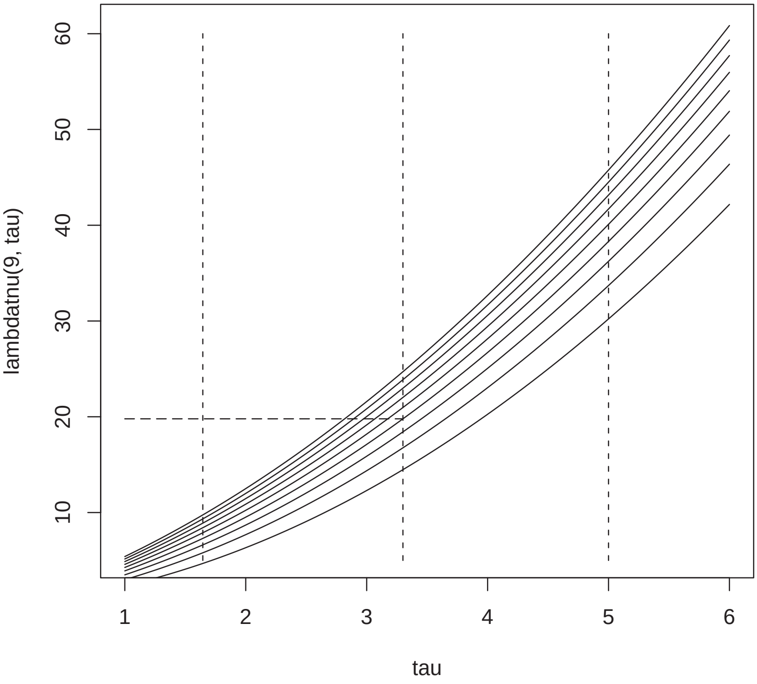


Figure 22.6 Plots of the graphs $(\tau, \lambda(\nu, \tau))$ defined in (22.10) for $\nu = 1, 2, \ldots, 9$ from the lowest line ($\nu = 1$) to the highest ($\nu = 9$). The vertical dashed lines correspond to weak, moderate and strong expected evidence ($\tau = 1.645$, 3.3 and 5, respectively). These plots allow for a quick determination of sample sizes required to achieve a desired expected evidence; see text for details.

22.4.2 Sample size required to obtain a desired power

Property K_2 of the Key Inferential Function in Section 17.2.2 described how the power $1-\beta = \Pi(\theta_1)$ against an alternative θ_1 at level α of a Neyman–Pearson test based on an evidence statistic T was related to the expected evidence $\tau = \mathrm{E}[T]$, namely $\tau = z_{1-\alpha} + z_{1-\beta}$. This statement is only true when the distribution of the evidence is exactly normal under both the null $\theta = 0$ and alternative $\theta = \theta_1$ hypotheses. For the chi-squared statistic the distribution of the evidence T_ν defined by (22.1) is very close to normal for alternatives but not so under the null hypothesis. Therefore for the *vst* of Definition 22.2 we define $h_{\nu,1-\alpha} = h_\nu(F_\nu^{-1}(1-\alpha))$, so that the exact $1-\alpha$ quantile of $T_\nu = h_\nu(S)$ is given by $h_{\nu,1-\alpha}$. Then a better approximation to the relationship between τ, level α and power $1-\beta$ is

$$\tau \;=\; h_{\nu,1-\alpha} + z_{1-\beta}. \tag{22.11}$$

This relationship, together with the methodology developed in Section 22.4.1, allows one to choose the sample size required to obtain power $1-\beta = \Pi(\theta_1) = P_{\theta_1}(T_\nu \geq h_{\nu,1-\alpha})$ against θ_1 at level α. One only needs to determine τ from (22.11), substitute it in $\lambda(\nu,\tau)$ of (22.10), and find N, the smallest integer greater than or equal to $\lambda(\nu,\tau)/\theta_1$. For example, with $\nu = K-1 = 4$, level $\alpha = 0.05$ and power $\beta = 0.8$, the relevant $\tau = 1.50 + 0.84 = 2.34$, so $\lambda(4, 2.34) = 11.9$. For alternative $\theta_1 = 1$, one requires a sample size of $N = 12$, while for alternative $\theta_1 = 0.2$, it is 60.

22.5 Evidence for $\lambda > \lambda_0$

The methods developed in Section 22.2 for testing $\lambda = 0$ versus $\lambda > 0$ can easily be extended to situations where one wants to test the hypotheses $\lambda \leq \lambda_0$ versus $\lambda > \lambda_0$, where the boundary λ_0 is positive.

We continue to use the notation N observations and the test statistic $S \sim \chi^2_\nu(\lambda)$. The parameter of interest is $\theta = \lambda/N$.

Definition 22.6 *Given the model* $S \sim \chi^2_\nu(\lambda)$, $\lambda \geq 0$, *and a fixed* $\lambda_0 > 0$. *The evidence for testing* $\lambda < \lambda_0$ *versus* $\lambda > \lambda_0$ *is defined by*

$$T_\nu(\lambda_0) \;=\; T_\nu - \sqrt{N}\,\mathcal{K}_{N,\nu}(\theta_0), \tag{22.12}$$

where T_ν *is defined in (22.1) and* $\mathcal{K}_{N,\nu}$ *is its associated Key Inferential Function (22.2). The magnitude of negative values of* $T_\nu(\lambda_0)$ *are positive evidence for* $\lambda < \lambda_0$ *while positive values are evidence for* $\lambda > \lambda_0$.

It follows from (22.5) that $\mathrm{E}[T_\nu(\lambda_0)] \doteq \sqrt{N}\{\mathcal{K}_{N,\nu}(\theta) - \mathcal{K}_{N,\nu}(\theta_0)\}$, *so the associated Key Inferential Function of* $T_\nu(\lambda_0)$ *is given for each* $\theta = \lambda/N$ *by*

$$\mathcal{K}_{N,\nu,\lambda_0}(\theta) = \mathcal{K}_{N,\nu}(\theta) - \mathcal{K}_{N,\nu}(\theta_0). \tag{22.13}$$

This $T_\nu(\lambda_0)$ inherits from T_ν properties E_1 to E_4 of Section 16.1.1 for a measure of evidence: monotonicity in the test statistic S, an expected evidence growing from 0 as the parameter increases from the null, a stabilized variance near 1 and approximate normality. As a special case, $T_\nu(0) = T_\nu$ and $\mathcal{K}_{N,\nu,0} = \mathcal{K}_{N,\nu}$.

22.6 Summary

In this chapter we derived a *vst* for any statistic S having a noncentral chi-squared distribution. This transformation required a smooth symmetrization about the null median so that the resulting evidence T_v is not only defined and variance stabilized, but also approximately normal for all values of S. It turns out that the Key Inferential Function for this model requires a bias correction term in order to be useful for inference. One can use this Key to derive upper confidence bounds and two-sided confidence intervals for the noncentrality parameter, and simulations demonstrate their accuracy providing the parameter is not too near zero. In addition, a slight modification of the Key enables one to carry out accurate sample size calculations to achieve a desired amount of evidence for an alternative of interest. Finally, we showed that the transformation is easily modified to allow one to find evidence for the noncentrality parameter exceeding a positive constant.

23

Evidence in *F*-tests

23.1 Variance stabilizing transformations for the noncentral *F*

The F-test is commonly used in the analysis of experiments in order to assess the importance of effects compared to the background noise level. In the one-way ANOVA with unequal sample sizes an outcome variable Y is observed under K different conditions, which may be different locations, different doses, or different treatments. This results in k samples $Y_{s1}, \dots, Y_{sn_s}$ for $(s = 1, \dots, K)$ with total sample size $N = n_1 + \cdots + n_K$. In the fixed effects model (FEM), the observations have expectation $\mathrm{E}(Y_{si}) = \mu_s$ and constant variance σ^2. The F-test statistic is

$$S = \frac{\sum_{s=1}^{K} n_s \left(\bar{Y}_s - \bar{Y}\right)^2 / (K-1)}{\sum_{s=1}^{K} \sum_{i=1}^{n_s} \left(Y_{si} - \bar{Y}_s\right)^2 / \sum_{s=1}^{K} (n_s - 1)}, \tag{23.1}$$

where $\bar{Y}_s$ is the mean of sample s and $\bar{Y}$ is the mean of all the observations. Denote the expected value of $\bar{Y}_{si}$ by μ_s and consider the null hypothesis $\mu_1 = \mu_2 = \cdots = \mu_K$. If this hypothesis is actually true and if the measurements have a normal distribution with constant variance and are independent of each other, then the test statistic S has an F-distribution with $\nu_n = K - 1$ and $\nu_d = \sum_{s=1}^{K}(n_s - 1) = N - K$ degrees of freedom for the numerator and the denominator, respectively. The proof of this results requires the use of linear transformations and knowledge of their effects on multivariate normal random vectors. It suffices to say that one can show that both the numerator and the denominator are proportional to chi-squared random variables, which furthermore are independent. The formal definition of the F-distribution is as follows.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

Definition 23.1 Let $U \sim \chi^2_{\nu_n}$ and $V \sim \chi^2_{\nu_d}$ be two independent *chi-squared random variables, with ν_n and ν_d degrees of freedom, respectively. The distribution of $S = (U/\nu_n)/(V/\nu_d)$ is said to be an F_{ν_n,ν_d} distribution.*

If an alternative holds, that is, if at least one of the expected values μ_s is different from the rest, then a noncentral F-distribution results. The behavior of the denominator of S is not affected by the fact that the null hypothesis is false. The numerator, however, changes to a noncentral chi-squared variable with the noncentrality parameter equal to

$$\lambda = \sum_{s=1}^{K} n_s(\mu_s - \mu)^2/\sigma^2, \tag{23.2}$$

where $\mu = \sum_{s=1}^{K} n_s\mu_s / \sum_{s=1}^{K} n_s$.

A nice overview of the issues and several proposals for variance stabilizing transformations are given in Laubscher (1960). Formally, our noncentral F-distributions are defined as follows.

Definition 23.2 *Let $U \sim \chi^2_{\nu_n}(\lambda)$ be a noncentral chi-squared random variable with noncentrality parameter $\lambda > 0$ and ν_n degrees of freedom. Let $V \sim \chi^2_{\nu_d}$ be an independent chi-squared random variable with ν_d degrees of freedom. The distribution of $S = (U/\nu_n)/(V/\nu_d)$ is said to be a noncentral F-distribution, $ncF_{\nu_n,\nu_d}(\lambda)$.*

The central F-distribution corresponds to $\lambda = 0$.

To derive the *vst* we have to express the variance of a noncentral F variable in terms of its expectation. These two quantities are

$$\begin{aligned}
\text{expectation} &= \nu_d(\nu_n + \lambda)/(\nu_n(\nu_d - 2)) \\
\text{variance} &= \frac{2\,\nu_d^2(\nu_n + \lambda)^2 + 2\,\nu_d^2(\nu_n + 2\lambda)(\nu_d - 2)}{\nu_n^2\,(\nu_d - 2)^2(\nu_d - 4)} \\
&= \frac{2}{\nu_d - 4}\left(\left(\text{expectation} + \frac{\nu_d}{\nu_n}\right)^2 - \frac{\nu_d^2\,(\nu_n + \nu_d - 2)}{\nu_n^2\,(\nu_d - 2)}\right).
\end{aligned}$$

The variance exists when $\nu_d > 4$ and the expectation is always larger than $\nu_d/(\nu_d - 2)$, which is the value of the expectation when $\lambda = 0$. As a consequence, the *vst* derived from the expression of the variance as a function of the expectation is not defined for all possible values of the statistic S. This variance stabilizing transformation $h(S)$ for a noncentral F variable with parameters ν_n, ν_d and λ satisfies

$$h'(x) = \sqrt{\frac{(\nu_d - 4)/2}{(x + \nu_d/\nu_n)^2 - c^2(\nu_n, \nu_d)}},$$

where the positive constant is $c^2(\nu_n, \nu_d) = \nu_d^2(\nu_n + \nu_d - 2)/(\nu_n^2(\nu_d - 2)) > 0$. The solution of this differential equation involves the hyperbolic cosine function, which is defined for any real number x by $\cosh(x) = (\mathrm{e}^x + \mathrm{e}^{-x})/2$. It is easy to verify that this function is symmetric. Only the positive branch, that is, $\cosh(x)$ for $x \geq 0$, is of interest to us. The inverse value x satisfies

$$\begin{aligned}(\mathrm{e}^x + \mathrm{e}^{-x}) = 2y &\Leftrightarrow (\mathrm{e}^x)^2 - 2y\mathrm{e}^x + 1 = 0\\ \Leftrightarrow \mathrm{e}^x = y \pm \sqrt{y^2 - 1} &\Leftrightarrow \cosh^{-1}(y) = \pm \ln\left(y + \sqrt{y^2 - 1}\right).\end{aligned}$$

The derivative of this function is exactly what we need for our *vst* . For the positive root we have

$$\frac{\mathrm{d}}{\mathrm{d}y}\cosh^{-1}(y) = \frac{1 + y/\sqrt{y^2 - 1}}{y + \sqrt{y^2 - 1}} = \frac{1}{\sqrt{y^2 - 1}}.$$

The *vst* thus is

$$\begin{aligned}h(S) &= \sqrt{(\nu_d - 4)/2}\,\cosh^{-1}\left(\frac{S + \nu_d/\nu_n}{c(\nu_n, \nu_d)}\right)\\ &= \sqrt{(\nu_d - 4)/2}\,\cosh^{-1}\left(\frac{\nu_n S + \nu_d}{\sqrt{\nu_d^2\,(\nu_n + \nu_d - 2)/(\nu_d - 2)}}\right).\end{aligned} \tag{23.3}$$

Strictly speaking, this is only valid for $S > \nu_d/(\nu_d - 2)$. For smaller values, one can still use it, but when the quotient inside $\cosh^{-1}$ becomes smaller than one, the corresponding value of h no longer exists. This difficulty is discussed by Laubscher (1960), who then switches to transformations in which the noncentrality parameter λ has to be estimated.

In order to extend the definition of the *vst* we will follow the general procedure outlined for the chi-squared test, that is, (1) re-center the function $h(S)$ such that it is equal to zero at the median $\mathrm{med}_{\nu_n,\nu_d}$ of the null distribution F_{ν_n,ν_d} and (2) flip the values for arguments above the median to those below the median in a symmetric fashion. Also in analogy to the chi-squared case, it is useful to modify the above function $h(S)$ and to bring the median into play in its definition.

The F-test turns into the chi-squared test when the number of degrees of freedom in the denominator is large. Consider the example of the one-way ANOVA (23.1) and suppose the sample sizes used in the experiment are fairly large, so that ν_d is large and ν_n/ν_d is small. We will now expand (23.3) in order to see how it compares to the *vst* considered in Chapter 22. The denominator of the argument of the inverse of the hyperbolic cosine function can be rewritten as

$$\begin{aligned}\nu_d^2\,(\nu_n + \nu_d - 2)/(\nu_d - 2) &= \nu_d^2\,(1 + (\nu_n - 2)/\nu_d)/(1 - 2/\nu_d)\\ &\doteq \nu_d^2\,(1 + (\nu_n - 2)/\nu_d)(1 + 2/\nu_d)\\ &\doteq \nu_d^2\,(1 + \nu_n/\nu_d).\end{aligned}$$

Note that the subtraction of 2 in the numerator and denominator cancels out in the limit. The inverse hyperbolic cosine function is evaluated at

$$\begin{aligned}(\nu_d + \nu_n\, S)/\left(\nu_d\,(1+\nu_n/\nu_d)^{1/2}\right) &\doteq (1 + (\nu_n\, S)/\nu_d)(1 - \nu_n/(2\,\nu_d))\\ &\doteq 1 + \frac{\nu_n\, S}{\nu_d} - \frac{\nu_n/2}{\nu_d}.\end{aligned}$$

When evaluating $\cosh^{-1}(1+\epsilon)$ for a small value of ϵ we find

$$\begin{aligned}\cosh^{-1}(1+\epsilon) &= \ln\left(1+\epsilon+\sqrt{(1+\epsilon)^2-1}\right)\\ &= \ln\left(1+\epsilon+\sqrt{2\epsilon+\epsilon^2}\right)\\ &\doteq \sqrt{2\epsilon}.\end{aligned}$$

It follows that for large ν_d and small ν_n/ν_d

$$\begin{aligned}h(S) &\doteq \sqrt{(\nu_d-4)/2}\,\sqrt{2\left(\frac{\nu_n\, S}{\nu_d} - \frac{\nu_n/2}{\nu_d}\right)}\\ &\doteq \sqrt{1-4/\nu_d}\,\sqrt{\nu_n\, S - \nu_n/2} \doteq \sqrt{\nu_n\, S - \nu_n/2}.\end{aligned}$$

When ν_d is large compared to ν_n, it follows that $\nu_n\, S$ is a noncentral $\chi^2_{\nu_n}$ variable, whereas the denominator of the F-test statistic (23.1) is approximately equal to the variance of the measurement error. We now compare the above expression with the evidence (22.1) for the chi-squared test Y. For values of Y larger than the median m_{ν_n} of the $\chi^2_{\nu_n}$ distribution, this evidence is up to the re-centering equal to

$$\sqrt{Y - \mathrm{m}_{\nu_n}/2} \approx \sqrt{\nu_n\, S - \mathrm{m}_{\nu_n}/2}.$$

This would be exactly equal to $h(S)$, if we replaced in $h(S)$ the half-mean $\nu_n/2$ by the half-median $\mathrm{m}_{\nu_n}/2$, which in turn is approximately equal to $\nu_n\,\mathrm{med}_{\nu_n,\nu_d}/2$. Looking back over the preceding development, we note that we could achieve the necessary change by replacing $(\nu_n + \nu_d - 2)$ by $(\nu_n\,\mathrm{med}_{\nu_n,\nu_d} + \nu_d - 2)$ inside the inverse hyperbolic cosine function. To further simplify the formula, we leave out the subtraction of 2 to arrive at

$$\sqrt{(\nu_d-4)/2}\,\cosh^{-1}\left(\frac{\nu_n\, S + \nu_d}{\sqrt{\nu_d^2\,(\nu_n\,\mathrm{med}_{\nu_n,\nu_d} + \nu_d)/\nu_d}}\right).$$

To center the transformation, we finally subtract the value at $S = \mathrm{med}_{\nu_n,\nu_d}$, which is equal to

$$\sqrt{(\nu_d-4)/2}\,\cosh^{-1}\left(\frac{\nu_n\,\mathrm{med}_{\nu_n,\nu_d} + \nu_d}{\sqrt{\nu_d\,(\nu_n\,\mathrm{med}_{\nu_n,\nu_d} + \nu_d)}}\right).$$

Simulations of this transformation for small numbers of degrees of freedom show that the multiplier $\sqrt{(\nu_d - 4)/2}$ leads to variances below the target value of 1. A further worthwhile modification consists in omitting the subtraction of 4 from ν_d.

This leads to the following definition of evidence.

Definition 23.3 *Let S be an F-test statistic of the null hypothesis $\lambda = 0$ versus $\lambda > 0$ with ν_n degrees of freedom for the numerator and ν_d degrees of freedom for the denominator. The corresponding evidence is defined as*

$$T = T(S) = \operatorname{sign}(S - \operatorname{med}_{\nu_n,\nu_d}) \sqrt{\frac{\nu_d}{2}} \left(\cosh^{-1}\left(\frac{\nu_n S^* + \nu_d}{\sqrt{\nu_d (\nu_n \operatorname{med}_{\nu_n,\nu_d} + \nu_d)}} \right) - \cosh^{-1}\left(\sqrt{\frac{\nu_n \operatorname{med}_{\nu_n,\nu_d} + \nu_d}{\nu_d}} \right) \right).$$

In this formula, S^* denotes the flipped value of the test statistic, equal to

$$S^* = \begin{cases} S, & \text{if } S \geq \operatorname{med}_{\nu_n,\nu_d} \\ F^{-1}_{\nu_n,\nu_d}\left(1 - F_{\nu_n,\nu_d}(S)\right), & \text{if } S < \operatorname{med}_{\nu_n,\nu_d}. \end{cases}$$

Recall that the inverse of the hyperbolic cosine function is

$$\cosh^{-1}(y) = \ln(y + \sqrt{y^2 - 1}).$$

Figure 23.1 shows the evidence as a function of the test statistic for two values of ν_d and increasing values of ν_n. Figure 23.2 illustrates the convergence to the chi-squared case when the number of degrees of freedom in the denominator grows.

23.2 The evidence distribution

The evidence defined in the previous section is a monotonic transformation $T(S)$ of the test statistic S, which itself has a noncentral F-distribution. It follows that the evidence $T(S)$ has density

$$f_{\nu_n,\nu_d,\lambda}(S(T))S'(T),$$

where $S(T)$ is the inverse transformation and $f_{\nu_n,\nu_d,\lambda}(S)$ is the noncentral F density. From Definition 23.3 it follows that

$$S^*(T) = \cosh\left(|T|\sqrt{\frac{2}{\nu_d}} + \cosh^{-1}\left(\sqrt{\frac{\nu_n \operatorname{med}_{\nu_n,\nu_d} + \nu_d}{\nu_d}} \right) \right) \times \frac{\sqrt{\nu_d (\nu_n \operatorname{med}_{\nu_n,\nu_d} + \nu_d)}}{\nu_n} - \frac{\nu_d}{\nu_n},$$

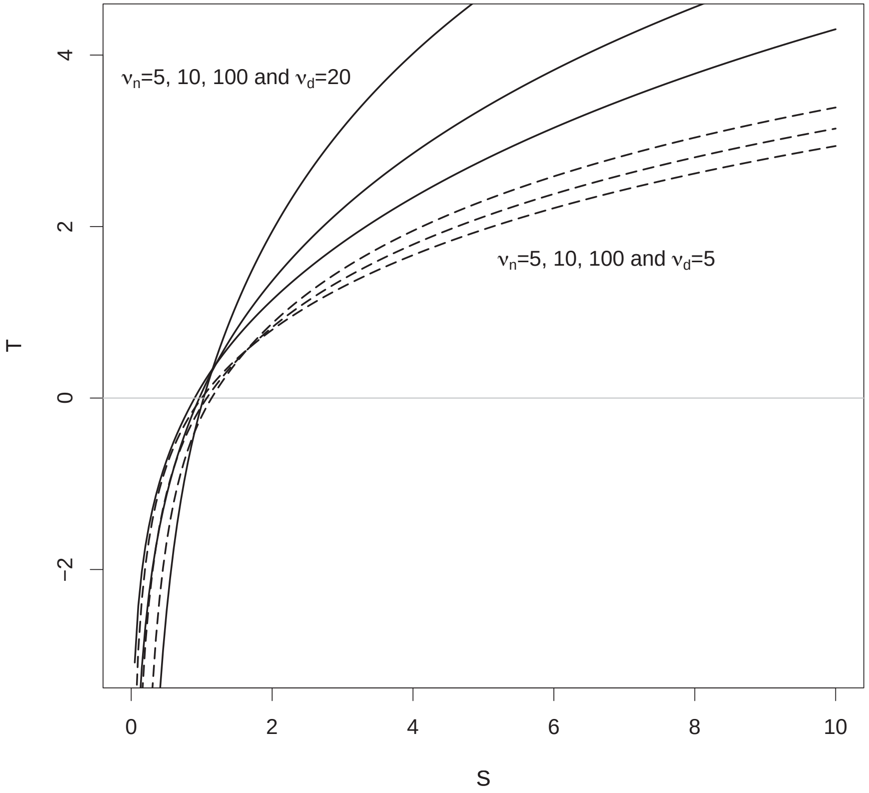


Figure 23.1 The curves show the evidence in an F-test statistic S for $\nu_n = 5,\ 10,\ 100$ and $\nu_d = 5,\ 20$.

whose derivative is

$$\frac{\mathrm{d}}{\mathrm{d}T}S^*(T) = \sinh\left(|T|\sqrt{\frac{2}{\nu_d}} + \cosh^{-1}\left(\sqrt{\frac{\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d}{\nu_d}}\right)\right)$$
$$\times \mathrm{sign}(T)\ \sqrt{2\nu_d(\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d)}/(\nu_n\sqrt{\nu_d}).$$

Note that $S \geq \mathrm{med}_{\,\nu_n,\nu_d} \Leftrightarrow T \geq 0 \Leftrightarrow S^*{=}S$. In this case, $S(T){=}S^*(T)$. For $T < 0$, we have $S^*(S){=}F^{-1}_{\nu_n,\nu_d}\left(1{-}F_{\nu_n,\nu_d}(S)\right)$, which has inverse $S(S^*){=}F^{-1}_{\nu_n,\nu_d}(1{-}F_{\nu_n,\nu_d}(S^*))$. Here, $S(T) = S(S^*(T))$. The final result we need for calculating the density of T is

$$\frac{\mathrm{d}}{\mathrm{d}S^*}\ S(S^*) = -f_{\nu_n,\nu_d}(S^*)/f_{\nu_n,\nu_d}(S).$$

The density of T is thus equal to

$$f(T) = \begin{cases} f_{\nu_n,\nu_d,\lambda}(S(T))(\mathrm{d}S^*/\mathrm{d}T)\,, & \text{if } \ T \geq 0 \\ f_{\nu_n,\nu_d,\lambda}(S(T))(\mathrm{d}S^*/\mathrm{d}T)\,(\mathrm{d}S/\mathrm{d}S^*)\,, & \text{if } \ T < 0. \end{cases} \tag{23.4}$$

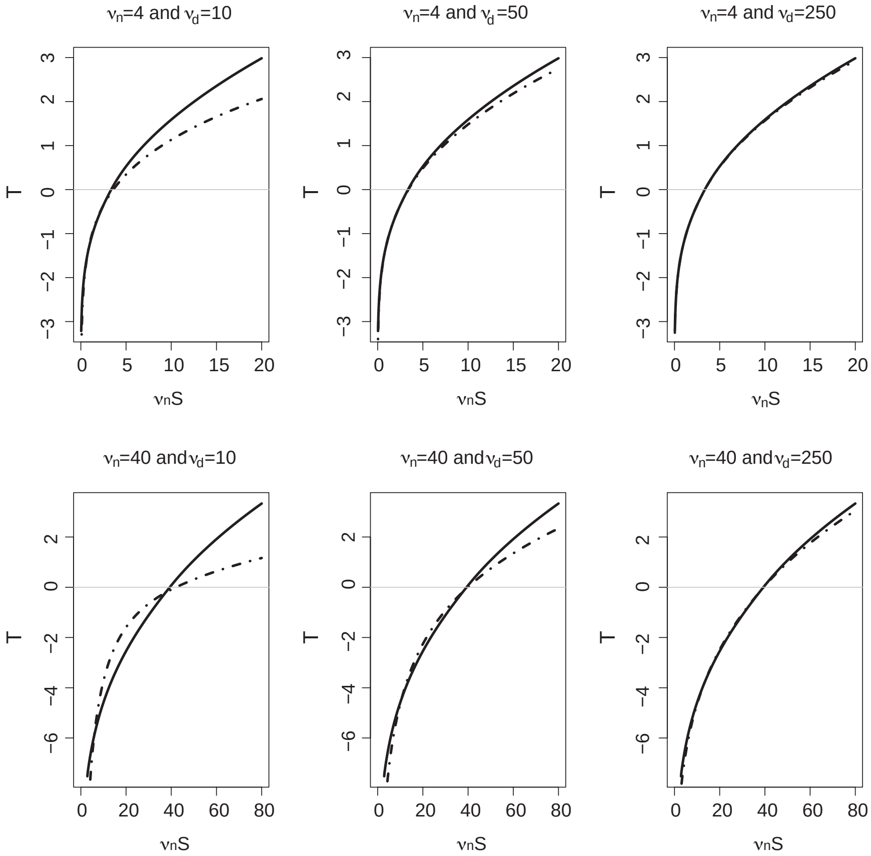


Figure 23.2 These plots show the evidence as a function of the test statistic for the chi-squared test (solid curves) and for the F-test (dashed curves). To make them comparable, the abscissa for the F-test is $\nu_n\, S$. The horizontal line indicates zero evidence.

When $\lambda = 0$, the density of the evidence T is symmetric around zero and equal to $f(T) = f_{\nu_n,\nu_d,\lambda=0}(S^*(T))(\mathrm{d}S^*/\mathrm{d}T)$. Figure 23.3 shows plots of these densities for a selection of values for degrees of freedom and noncentrality parameter. The agreement between the normal density and the density of the evidence is on the whole quite good. The biggest discrepancies occur for $\lambda = 0$, which is the null hypothesis being tested. For small ν_n (top row in Figure 23.3) the variance of the evidence is visibly smaller than 1. When ν_d is smaller than ν_n, the density of the evidence has a slightly increased variance. In the top row, when $\lambda = 0$, it is evident that the derivative of the density of T is not smooth at 0. This is due to a discontinuity of the derivative of the transformation $\mathrm{d}/\mathrm{d}S\ T(S)$ at the point $S = \mathrm{med}_{\nu_n,\nu_d}$. In the other plots, the lack of smoothness at 0 is less visible, but it is still present.

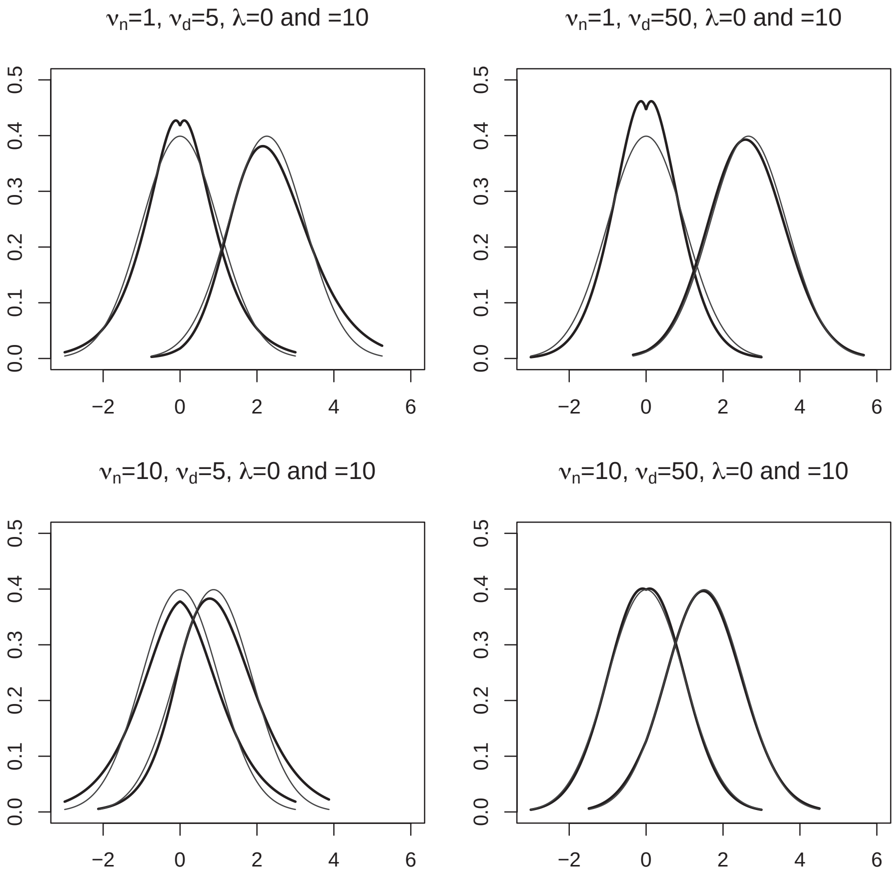


Figure 23.3 The four panels show the densities of the evidence for four couples of degrees of freedom. For comparison, the standard normal density φ is included (thin line).

23.3 The Key Inferential Function

The evidence T has expected value $E(T) \doteq \sqrt{N}\,\mathcal{K}(\nu_n, \nu_d, \lambda)$, where N is the sample size, that is, the total number of observations. A first approximation for this expectation is obtained by calculating the evidence we would obtain with the mean value of the test statistic S:

$$\begin{aligned} E(T) &\doteq T(S = \nu_d\,(\nu_n+\lambda)/(\nu_n(\nu_d-2))) \\ &\doteq \sqrt{\frac{\nu_d}{2}}\left(\cosh^{-1}\left(\frac{(\nu_n+\lambda)/(1-2/\nu_d)+\nu_d}{\sqrt{\nu_d\,(\nu_n \mathrm{med}_{\,\nu_n,\nu_d}+\nu_d)}}\right)\right. \\ &\qquad\qquad \left. -\cosh^{-1}\left(\sqrt{\frac{\nu_n \mathrm{med}_{\,\nu_n,\nu_d}+\nu_d}{\nu_d}}\right)\right). \end{aligned}$$

As shown in Figure 23.1, the transformation $T(S)$ to evidence is concave and the above approximation is an upper bound on the actual expectation. This can be seen very clearly when $\lambda = 0$. We saw above that the distribution of T is symmetric around zero, but obviously the above approximate expectation is not zero. We could now go ahead and compute a correction based on the second derivative of the transformation $T(S)$. However, the case when $\lambda = 0$ suggests a simpler remedy, namely to use the following formula:

$$E(T) \doteq \sqrt{\frac{\nu_d}{2}} \left(\cosh^{-1} \left(\frac{\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \lambda + \nu_d}{\sqrt{\nu_d \, (\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d)}} \right) \right. \\ \left. - \cosh^{-1} \left(\sqrt{\frac{\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d}{\nu_d}} \right) \right), \tag{23.5}$$

which clearly is equal to zero when $\lambda = 0$. The matching normal distributions in Figure 23.3 are all centered at the approximate expected value calculated by the above expression and we note that it does a very good job.

The Key Inferential Function corresponding to (23.5) is as follows.

Definition 23.4 *The Key Inferential Function when using an F-test is equal to*

$$\mathcal{K}(\lambda) = \sqrt{\frac{\nu_d}{2N}} \left(\cosh^{-1} \left(\frac{\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \lambda + \nu_d}{\sqrt{\nu_d \, (\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d)}} \right) \right. \\ \left. - \cosh^{-1} \left(\sqrt{\frac{\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d}{\nu_d}} \right) \right).$$

As in all the other tests discussed in this book, the Key Inferential Function translates the apparent effect λ (see, for example, Equaton (23.2)) into a statistically meaningful effect size. Figure 23.4 shows how the key varies with the noncentrality parameter. For λ close to zero one has $\mathcal{K}(\lambda) \doteq \mathcal{K}'(0)\,\lambda$ with the derivative approximately equal to

$$\mathcal{K}'(0) \approx \sqrt{\frac{1 - \nu_n/N}{2\nu_n}} \, \frac{1}{\sqrt{N}}.$$

Clearly, the larger ν_n, the smaller the rate of increase. For larger values of the noncentrality parameter, the Key Inferential Function grows logarithmically.

The interval $T \pm 1.96$ is a confidence interval for $\sqrt{N}\,\mathcal{K}$, which can be inverted to obtain a confidence interval for the noncentrality parameter λ. The inverse function $\lambda(\mathcal{K})$ is

$$\lambda(\mathcal{K}) \; = \cosh \left(\sqrt{N}\mathcal{K} \, \sqrt{2/\nu_d} + \cosh^{-1} \left(\sqrt{\frac{\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d}{\nu_d}} \right) \right) \\ \times \sqrt{\nu_d \, (\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d)} - (\nu_n \mathrm{med}_{\,\nu_n,\nu_d} + \nu_d).$$

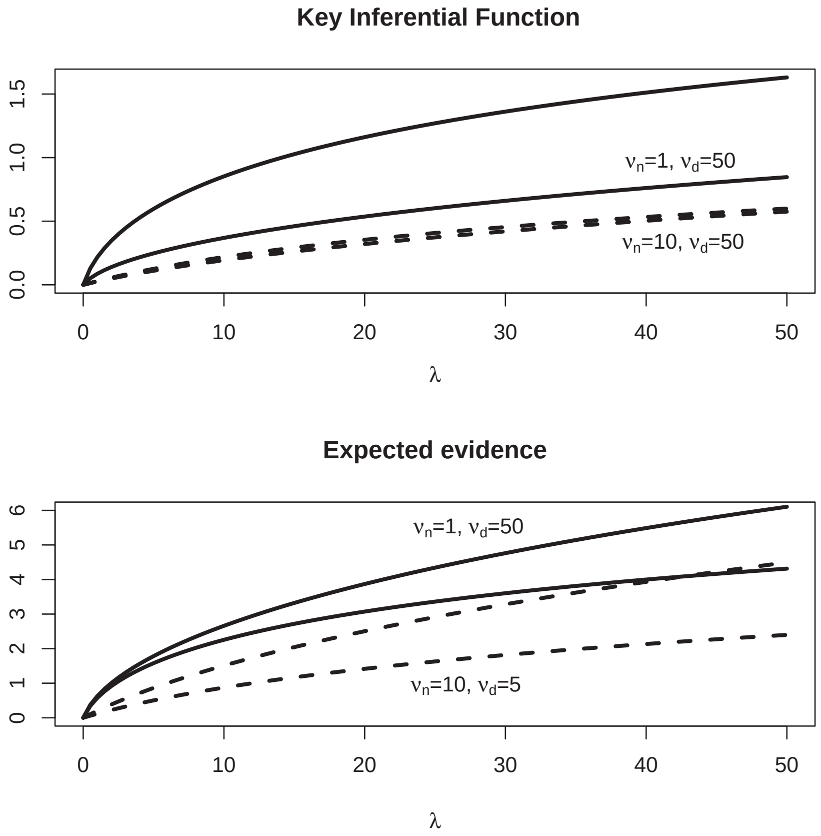


Figure 23.4 The upper panel shows the Key Inferential Function for the same combination of degrees of freedom used in Figure 23.3, namely all combinations of $\nu_n = 1$(solid) or 10 (dashed) and $\nu_d = 5$ or 50. The one-way ANOVA setting, where $N = \nu_n + \nu_d + 1$, is assumed. The lower panel shows the total expected evidence from the experiment.

To compute the desired confidence limits, one has to substitute $T \pm 1.96$ for $\sqrt{N}\mathcal{K}$.

Transforming the test statistic S to the evidence T not only provides a calibrated scale on which to judge the outcome of a statistical test and to combine test results, but it also makes sample size and power calculations easy. In testing $\mu = \mu_0$ versus $\mu = \mu_0 + \Delta > \mu_0$ with a single observation of a normal random variable with unit variance, the true discovery rate or power at the alternative $\Delta > 0$ is

$$P(T > z_{1-\alpha}) = P(T - \Delta > z(1-\alpha) - \Delta) = \Phi(z_{1-\alpha} - \Delta),$$

where Φ denotes the unit normal distribution function, $z_{1-\alpha}$ is its $1-\alpha$ quantile and α is the probability of a false discovery or the type-I error rate. It follows that in order to reach a probability for a true discovery or power of $1 - \beta$, the alternative must satisfy

$$z_{1-\alpha} - \Delta = z_\beta = -z_{1-\beta} \Leftrightarrow \Delta = z_{1-\alpha} + z_{1-\beta}.$$

After transformation to evidence, we are approximately in the situation of a test as described above, with the Key Inferential Function in the role of the shift parameter,

$$\Delta = \sqrt{N}\mathcal{K}(\lambda).$$

The power function of an F-test is thus approximately equal to

$$\text{Power}(\lambda) = \Phi(z_{1-\alpha} - \sqrt{N}\mathcal{K}(\lambda)).$$

In a typical situation, the noncentrality parameter λ is approximately known or it is assumed to have a certain size. The number of degrees of freedom in the numerator ν_n is also typically known, whereas the number of degrees of freedom in the denominator ν_d and with it the total size N of the study can be adjusted. In order to reach power $1 - \beta$, the total sample size has to be chosen as

$$N = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\mathcal{K}(\lambda)}\right)^2.$$

To illustrate the approximation, Figure 23.5 shows the power curves as a function of the noncentrality parameter for different sample sizes.

23.3.1 Refinements

The Key Inferential Function and the normal approximation

$$T \sim \text{Normal}(\mu = \sqrt{N}\mathcal{K}(\lambda), \sigma^2 = 1)$$

lead to satisfactory results. Its main advantage is the simplicity of its use. No complex approximations need to be calculated. Knowledge of the *vst*, which is a simple function given in closed form, is sufficient. However, all the elements in the above distributional approximation contains errors. The distribution of the evidence T is not exactly normal, its expectation is not exactly equal to the μ indicated and its variance is not exactly equal to 1. One could try to improve the approximate formulas given, but this would lead to considerably more complicated expressions, which is reason enough for us not to pursue these ideas any further.

23.4 The random effects model

The standard F-test in one-way ANOVA (23.1) is based on the two sums of squares

$$\text{SS}_A = \sum_{k=1}^{K} n_k \left(\bar{Y}_k - \bar{Y}\right)^2,$$

and

$$\text{SS}_e = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \left(Y_{ki} - \bar{Y}_k\right)^2.$$

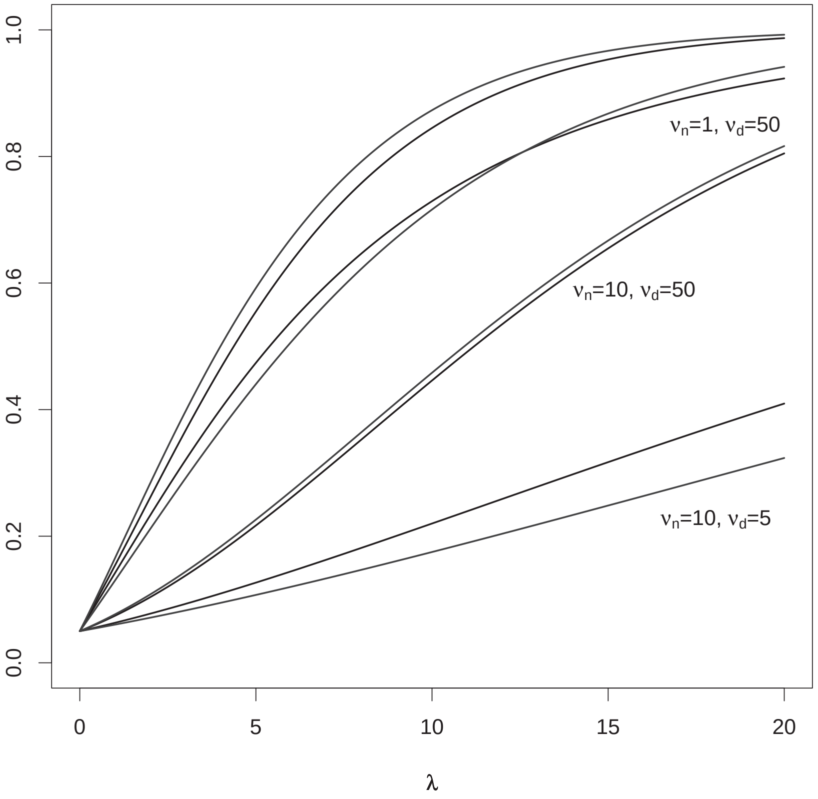


Figure 23.5 The approximate and true power of the F-test as a function of the noncentrality parameter λ. The values for the degrees of freedom are all combinations of $\nu_n = 1$ or 10 and $\nu_d = 5$ or 50. In each case, the quality of the approximation is satisfactory, with the exception of the couple $\nu_n = 5$ and $\nu_d = 5$, where the power is overestimated when using the key inferential statistic.

Under the fixed effects model (FEM) $Y_{ki} = \mu_k + e_{ki}$ with iid random variables $e_{ki} \sim N(0, \sigma_e^2)$. Under the random effects model (REM) $Y_{ki} = \mu + a_k + e_{ki}$ with iid $a_k \sim N(0, \sigma_A^2)$ independent from the errors $e_{ki} \sim N(0, \sigma_e^2)$. The null hypothesis under the FEM is $H_0 : \mu_k = \mu$, and under the REM it is $H_0 : \sigma_A^2 = 0$.

Note that under the REM $\bar{Y}_k = a_k + \bar{e}_k \sim N(0, \sigma_A^2 + \sigma_e^2/n_k)$. Let $w_k = \sigma_A^2 + \sigma_e^2/n_k = \sigma_e^2(\theta + 1/n_k)$, and $W = \mathrm{diag}(w_k)$ the $K \times K$ diagonal matrix with w_k on the diagonal. Note that the parameter $\theta = \sigma_A^2/\sigma_e^2$ measures the standardized distance to the null hypothesis. The random vector $\mathbf{Z} = (\bar{Y}_1, \ldots, \bar{Y}_K)$ has distribution $N(0, W)$ and can be used to rewrite the sum of squares SS_A as

$$\mathrm{SS}_A = \boldsymbol{Z}^t \mathrm{B} \boldsymbol{Z}, \text{ with } \mathrm{B} = (b_{kt}) = n_k \left(\delta_{kt} - \frac{n_t}{N} \right).$$

In this last formula, δ_{kt} denotes Kronecker's delta, which is 1 if $k = t$ and 0 otherwise. For a balanced design, $n_k = n = N/K$, and $b_{kt} = n(\delta_{kt} - 1/K)$. Thus $\mathrm{SS}_A \sim (n\sigma_A^2 + \sigma_e^2)\chi^2_{K-1}$.

For both REM and FEM, $\mathrm{SS}_e \sim \sigma_e^2 \chi^2_{N-K}$, independently from SS_A. Thus for a balanced case

$$S = \frac{\mathrm{SS}_A/(K-1)}{\mathrm{SS}_e/(N-K)} \equiv \left(1 + n\,\frac{\sigma_A^2}{\sigma_e^2}\right) X = (1 + n\theta)X\,,$$

where $X \sim F_{\nu_n = K-1, \nu_d = N-K}$. This is different from the noncentral F-distribution we found previously for FEM. In REM the distribution under alternatives is a scaled central F-distribution.

23.4.1 Expected evidence in the balanced case

Using the formulas for the expectation and variance of a non-central F-distribution we find

$$\begin{aligned} \mathrm{E}(S) = \mu(\theta) &= (1 + n\theta)\frac{\nu_d}{\nu_d - 2} = (1 + n\theta)\frac{N-K}{N-K-2}\,, \\ \mathrm{Var}(S) = \sigma^2(\theta) &= \big(1 + n\theta\big)^2 \frac{2\nu_d^2(\nu_n + \nu_d - 2)}{\nu_n(\nu_d - 2)^2(\nu_d - 4)} \\ &= \big(1 + n\theta\big)^2 \frac{2(N-K)^2(N-3)}{(K-1)(N-K-2)^2(N-K-4)} \\ &= \mathrm{E}(S)^2 \frac{2(N-3)}{(K-1)(N-K-4)} = a\,\mu(\theta)^2\,, \end{aligned} \tag{23.6}$$

where $\theta = \sigma_A^2/\sigma_e^2$. The *vst* corresponding to (23.6) is $h(S) = a^{-1/2}\ln(S)$. Under the null hypothesis, $\theta = 0$ and $\mathrm{E}(h(S)) \doteq h(\mu(0)) + h''(\mu(0))\sigma^2(0)/2 = a^{-1/2}\ln(\mu(0))$ $-a^{1/2}/2$, and subtracting this term, we obtain the evidence

$$T = a^{-1/2}\ln(S/\mu(0)) + a^{1/2}/2. \tag{23.7}$$

The expected evidence under a general alternative is

$$\begin{aligned} \mathrm{E}(T) &= a^{-1/2}\ln(\mu(\theta)/\mu(0)) - a^{1/2}/2 \\ &= a^{-1/2}\ln(1 + n\theta) = \left(\frac{(K-1)(N-K-4)}{2(N-3)}\right)^{1/2} \ln(1 + n\theta) \\ &\doteq (K-1)^{1/2}[(n-1)/(2n)]\ln(1 + n\theta). \end{aligned} \tag{23.8}$$

In the last formula we made use of $N = Kn$, $N - K - 4 = K(n-1) - 4 \approx K(n-1)$ and $N - 3 \approx Kn$, which shows that this approximation holds when at least one of K or n is large.

23.4.2 Comparing evidence in REM and FEM

Under REM, the expected evidence is of order $K^{1/2}$ but depends on n at a much lower rate of $\ln(n)$. This is natural since one can improve the estimate of σ_A^2 only by increasing the number of samples K (and not the number of observations n in each). The noncentrality parameter λ is related to θ by the approximate equality $N\theta = \lambda$, which equals $nK\theta$ in the balanced case. Depending on the choice of K and n, the evidence in a REM may be bigger or smaller than in the corresponding FEM. Few (small K), but large (n large) samples is a design that is unfavorable to REM and favorable to FEM. When K and n are about equal (both around $\sqrt{N}$) then FEM tends to produce more evidence in favor of an alternative λ.

23.5 Summary

In this chapter we considered test statistics with an F-distribution. Under the null hypothesis, the test statistic is a ratio of independent chi-squared variables, each divided by its number of degrees of freedom. It often happens that under the alternatives the numerator becomes a noncentral chi-squared variable and it is for this case that we derived a *vst* . The evidences resulting from F-tests in fixed effects ANOVA models can be computed with the help of this transformation. Another large area of applications are regression models.

24

Evidence in Cochran's Q for heterogeneity of effects

Given K studies measuring potentially different effects μ_k for $k = 1, \ldots, K$ it is customary to test the null hypothesis of equal effects, or *homogeneity*, using the Q statistic introduced by Cochran (1954); it is a weighted sum of squared deviations of the effects from their weighted mean, and the topic of Section 24.1. The alternative hypothesis of *heterogeneity* asserts that $\mu_j \neq \mu_k$ for some $j \neq k$. Assuming the μ_k's can be estimated by asymptotically normal statistics, Q has, under the alternative of heterogeneity, a limiting noncentral chi-squared distribution, as shown in Section 24.1.1. Unfortunately, when the weights in Q need to be estimated, the distribution of Q often converges slowly to its limit, making p-values based on this limit of dubious value. Welch (1951) and James (1951) suggested a better moderate sample size approximation to the null distribution of Q which leads to the Welch F-test for homogeneity, and Kulinskaya and Staudte (2007) proposed an approximation to the distribution of Q under alternatives. But here we advocate thinking in terms of evidence for the alternative of heterogeneity. A *vst* of Q from Chapter 22 will find the large-sample evidence T_Q for heterogeneity of effects (see Section 24.1.2).

We also introduce another approach which makes Q useful for even moderately small sample sizes. The idea is to find the evidence in each of K studies, use such evidence to estimate a transformation of the standardized effect and then apply Cochran's Q *with known weights* to these transformed effects. The resulting Q^* has an approximate noncentral chi-squared distribution, and so is readily transformed into evidence on the canonical scale using the results of Chapter 22. The simple theory

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

and applications of this approach are the content of Section 24.1.3. In Section 24.2 the performances of Q and Q^* are examined through simulation studies.

The random effects model is discussed in Section 24.3. It is well known that Cochran's Q has the same null distribution under the fixed and random effects models. However, the alternative hypotheses are different for the two models, and for fixed K the evidence for the alternative of heterogeneity in Cochran's Q for the random effects model is smaller than for the fixed effects model.

24.1 Cochran's Q: the fixed effects model

It is customary in the meta-analytic literature to carry out a chi-squared test of the hypothesis of homogeneity of effects using Cochran's Q. If the test fails to reject it is then assumed the effects are equal, and if it does reject then an alternative model which allows for different effects is assumed. In this section we propose to measure the evidence for the alternative of unequal effects; that is, to replace the all-or-nothing approach of testing a null hypothesis with a measure of the evidence in the test statistic Q; the researcher then has more information with which to make a decision regarding the choice of models.

24.1.1 Background material

Assume the estimated effects $\hat{\mu}_k, k = 1, \ldots, K$ for the respective studies are mutually independent and satisfy $\sqrt{w_k}\,(\hat{\mu}_k - \mu_k) \to N(0, 1)$. Cochran's Q is defined by

$$Q = \sum_k \hat{w}_k (\hat{\mu}_k - \hat{\bar{\mu}}_{\hat{w}})^2, \tag{24.1}$$

where $\bar{\mu}_w = \sum w_k \mu_k / \sum_k w_k$ is the weighted effect, and $\hat{w}_k^{-1}$ is the estimated asymptotic variance w_k^{-1} of $\hat{\mu}_k$. We restrict attention to situations where for each k there are n_k observations in the kth study and $w_k^{-1} = \sigma_k^2 / n_k$ for a fixed, but usually unknown $\sigma_k^2 > 0$. In particular when the observations in the kth study are modeled by the normal distribution $N(\mu_k, \sigma_k^2)$, one estimates μ_k by the sample mean $\hat{\mu}_k = \bar{X}_k$ and w_k by $\hat{w}_k = n_k / s_k^2$, where s_k^2 is the sample variance.

Proposition 24.1 *Assume* $\boldsymbol{Y}$ *is a mutivariate* K*-vector with* $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \Sigma)$, *where* Σ *is a known non-singular diagonal matrix with inverse* $W = \Sigma^{-1}$. *Denote the kth diagonal element of* W *by* w_k *and define* $p_k = w_k / (\sum_j w_j)$, $\mu_w = \sum_k p_k \mu_k$, *and* $\bar{Y}_w = \sum_k p_k Y_k$. *Then the statistic* S *has a noncentral chi-squared distribution:*

$$S = \sum_k w_k (Y_k - \bar{Y}_w)^2 \sim \chi^2_{K-1}(\lambda), \tag{24.2}$$

where

$$\lambda = \sum_k w_k (\mu_k - \bar{\mu}_w)^2. \tag{24.3}$$

Proof: *Let P be the diagonal matrix with kth element p_k, I the $K \times K$ identity matrix and J the $K \times K$ matrix of 1's. Further introduce the symmetric matrix $C = (I - JP)'W(I - JP)$. Then it is easily checked that the quadratic form $\boldsymbol{Y}'C\boldsymbol{Y} = \sum_k w_k(Y_k - \bar{Y}_w)^2 = S$.*

It now follows from a standard result (Serfling (1980), pp. 128–129) that if $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \Sigma)$ with Σ nonsingular and if C is symmetric, then $\boldsymbol{Y}'C\boldsymbol{Y}$ has a noncentral chi-squared distribution if and only if $\Sigma\, C\, \Sigma\, C\, \Sigma = \Sigma\, C\, \Sigma$, in which case the degrees of freedom is the trace $\mathrm{tr}\,(C\Sigma)$ *and the noncentrality parameter is $\boldsymbol{\mu}'C\boldsymbol{\mu}$.*

To complete the proof, check that $C\Sigma = I - PJ$ and $JPJ = J$, so $\Sigma\, C\, \Sigma = \Sigma - J/tr(W) = \Sigma\, C\, \Sigma\, C\, \Sigma$. The degrees of freedom in the noncentral chi-squared distribution of S are $\mathrm{tr}\,(C\Sigma) = \mathrm{tr}\,(I - PJ) = K - 1$, *and the noncentrality parameter is $\lambda = \boldsymbol{\mu}'C\boldsymbol{\mu} = \sum_k w_k(\mu_k - \mu_w)^2$. Clearly this distribution is the central chi-squared distribution if and only if all μ_k are equal.*

It will be rare that the conditions of the above proposition are satisfied; rather it is usually tacitly assumed that the statistical model of interest is imbedded in a sequence of models indexed by a superscript $^{(i)}$, say, for which the statistic of interest, in our case Cochran's $Q^{(i)}$, has a limiting noncentral chi-squared distribution, and then this limiting distribution is used to approximate the unknown actual distribution of the particular ith model in hand. Therefore we consider some extensions of Proposition 24.1.

Proposition 24.2 *Fix the number of groups at K and define limiting sample proportions $\boldsymbol{r} = (r_1, \ldots, r_K)'$, all $r_k > 0$. Let $\boldsymbol{n}^{(i)} = (n_1^{(i)}, \ldots, n_K^{(i)})$, with total denoted $N^{(i)} = \sum_k n_k^{(i)}$, define a sequence of sample sizes for the K groups satisfying $\boldsymbol{n}^{(i)}/N^{(i)} \to \boldsymbol{r}$ as $i \to \infty$. For each $k = 1, \ldots K$ and every i let $\hat{\mu}_k^{(i)}$ be an estimator of μ_k based on the available $n_k^{(i)}$ observations, and assume $\{n_k^{(i)}\}^{1/2}\{\hat{\mu}_k^{(i)} - \mu_k\}/\sigma_k \to N(0, 1)$ in distribution as $i \to \infty$. Further assume the $\hat{\mu}_k$'s are mutually independent.*

In vector notation, for $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]'$ and $\hat{\boldsymbol{\mu}}^{(i)} = [\hat{\mu}_1^{(i)}, \ldots, \hat{\mu}_K^{(i)}]'$, it follows that

$$\boldsymbol{Y}^{(i)} = \left\{N^{(i)}\right\}^{1/2}\left\{\hat{\boldsymbol{\mu}}^{(i)} - \boldsymbol{\mu}\right\} \to \boldsymbol{Y} \sim N(\boldsymbol{0}, \Sigma) \tag{24.4}$$

in distribution as $i \to \infty$. Here Σ is a diagonal matrix with $\Sigma_{kk} = \sigma_k^2/r_k > 0$.

With these preliminaries, we may now find:

1. *Limiting distribution of $Q^{(i)} = \sum_k (n_k/\sigma_k^2)\left\{\hat{\mu}_k^{(i)} - \hat{\bar{\mu}}_w\right\}^2$ for unknown weights under the null hypothesis of homogeneity. For each i define $\{W^{(i)}\}^{-1}{=}\Sigma^{(i)}{=}$ Cov$\,[\boldsymbol{Y}^{(i)}]$, $P^{(i)}{=}W^{(i)}/\mathrm{tr}\,(W^{(i)})$ and $C^{(i)}{=}(I - JP^{(i)})'W^{(i)}(I{-}JP^{(i)})$. Under the hypothesis of homogeneity $\boldsymbol{\mu} = \mu\mathbf{1}_K$, we have*

$$Q^{(i)} = (\boldsymbol{\mu}^{(i)})'C^{(i)}\boldsymbol{\mu}^{(i)} = \sum_k \frac{n_k^{(i)}}{\sigma_k^2}\left\{\hat{\mu}_k^{(i)} - \hat{\bar{\mu}}_w\right\}^2 \to \chi^2_{K-1} \tag{24.5}$$

in distribution as $i \to \infty$.

Proof: *Observe that* $W^{(i)}/N^{(i)} \to \Sigma^{-1}$ *arising in Equation (24.4), so* $C^{(i)}/N^{(i)} \to D = (I - JP)'\Sigma^{-1}(I - JP)$. *Letting* $a_N = \mu\{N^{(i)}\}^{1/2}$ *and using* $\boldsymbol{Y}^{(i)} = \left\{N^{(i)}\right\}^{1/2}\left\{\hat{\boldsymbol{\mu}}^{(i)} - \boldsymbol{\mu}\right\}$, *we may write Equation (24.5) as*

$$Q^{(i)} = \left(\boldsymbol{Y}^{(i)} + a_N\mathbf{1}_K\right)' \frac{C^{(i)}}{N^{(i)}} \left(\boldsymbol{Y}^{(i)} + a_N\mathbf{1}_K\right) \qquad (24.6)$$

$$= \left(\boldsymbol{Y}^{(i)}\right)' \frac{C^{(i)}}{N^{(i)}} \boldsymbol{Y}^{(i)}$$

because $(I - JP^{(i)})\, a_N\mathbf{1}_K = \mathbf{0}_K$. *It now follows from Proposition 24.1 and the continuity theorem that* $Q^{(i)} \to \boldsymbol{Y}'D\boldsymbol{Y} \sim \chi^2_{K-1}$ *as* $i \to \infty$.

2. Limiting distribution of $Q^{(i)}$ for known weights under the alternative hypothesis of heterogeneity. *Assume for each i, the effect that was fixed at* μ_k *is replaced by* $\mu_k^{(i)} = \mu + \Delta_k/\{N^{(}i)\}^{1/2}$, *and that for each k,* $\{n_k^{(i)}\}^{1/2}\{\hat{\mu}_k^{(i)} - \mu_k^{(i)}\}/\sigma_k \to N(\eta_k, 1)$ *in distribution as* $i \to \infty$. *Here* $\eta_k = \Delta_k\sqrt{r_k}\,/\sigma_k$. *Then (24.4) is replaced by* $\boldsymbol{Y}^{(i)} = \left\{N^{(i)}\right\}^{1/2}\left\{\hat{\boldsymbol{\mu}}^{(i)} - \boldsymbol{\mu}^{(i)}\right\} \to \boldsymbol{Y} \sim N(\boldsymbol{\Delta}, \Sigma)$ *in distribution. Hence by Proposition 24.1 and the continuity theorem,*

$$Q^{(i)} \to \boldsymbol{Y}'D\boldsymbol{Y} \sim \chi^2_{K-1}(\lambda), \qquad (24.7)$$

where

$$\lambda = \boldsymbol{\Delta}'D\boldsymbol{\Delta} = \sum_k \frac{r_k}{\sigma_k^2}(\Delta_k - \bar{\Delta}_w)^2. \qquad (24.8)$$

Note that the noncentrality parameter in the ith problem is approximately $\lambda^{(i)} = N^{(i)}\theta^{(i)}$, *where*

$$\theta^{(i)} = \sum_k \frac{n_k^{(i)}}{\sigma_k^2}\left(\mu_k^{(i)} - \bar{\mu}_w^{(i)}\right)^2. \qquad (24.9)$$

3. Limiting distribution of $Q^{(i)}$ for unknown weights. *Suppose that the weights* $w_k^{(i)} = n_k^{(i)}/\sigma_k^2$ *are unknown, but there exists for each k estimators* $\hat{\sigma}_k^{(i)}$ *based on the available* $n_k^{(i)}$ *observations with* $\hat{\sigma}_k^{(i)} \to \sigma_k > 0$ *in probability. Then again by the continuity theorem,* $Q^{(i)}$ *has a limiting distribution given in the previous two parts of this proposition.*

24.1.2 Evidence for heterogeneity of fixed effects

Cochran's Q as defined in (24.1) is the standard test statistic for testing against homogeneity of the effects μ_k. Whenever Q has an approximate noncentral chi-squared distribution one can calibrate the evidence in it for heterogeneity of effects using the *vst* derived in Chapter 22.

Definition 24.1 *Let* $Q \sim \chi^2_\nu(\lambda)$, *with* $\nu = K - 1$ *and* $\lambda = \sum_k (n_k/\sigma_k^2)(\mu_k - \bar{\mu}_w)^2$. *The evidence in Q for heterogeneity is defined to be the* vst *in (22.1) applied to Q,* *and denoted* $T_Q = h_{K-1}(Q)$. *Then* $T_Q \sim N(\tau_Q, 1)$, *where the mean evidence is given by* $\tau_Q \doteq \sqrt{N}\,\mathcal{K}(\theta)$, *with the Key defined by (22.2) and* $\theta = \lambda/N$.

We consider some examples illustrating the performance of T_Q with various configurations of parameters in Section 24.2.

24.1.3 Evidence for heterogeneity of transformed effects

We have found in earlier chapters that an appropriate *vst* of a test statistic often leads to evidence T having an approximate $N(\tau, 1)$ distribution, where $\tau = \sqrt{N}\,\mathcal{K}(\delta)$, N is the sample size and δ is a standardized effect. For example, the $\sqrt{2n}\ \sinh^{-1}(t_{n-1}/\sqrt{2n}\,)$ transformation of the Student t-statistic (20.4) leads to $\tau \doteq \sqrt{n}\,\mathcal{K}(\delta)$, where $\delta = (\mu - \mu_0)/\sigma$.

Suppose now there are K studies, each entailing a standardized effect δ_k of interest, and one wants to know if the standardized effects are equal or not. One could apply the method of Cochran to the estimated effects $\hat{\delta}_k$, for $k = 1, \ldots, K$, provided one could find weights v_k for which $\sqrt{v_k}\,(\hat{\delta}_k - \delta) \to N(0, 1)$. Another approach is based on the existence of a *vst* of $\hat{\delta}_k$, as follows: let $T_k = h(\hat{\delta}_k) \sim N(\sqrt{n_k}\,\mathcal{K}(\delta_k), 1)$, at least approximately. Then the transformed effects $\kappa_k = \mathcal{K}(\delta_k)$ can be estimated by $\hat{\kappa}_k = T_k/\sqrt{n_k} \sim N(\kappa_k, 1/n_k)$. Hence Cochran's Q statistic can be calculated for the $\hat{\kappa}_k$'s with known weights n_k:

$$Q^* = \sum_k n_k(\hat{\kappa}_k - \hat{\bar{\kappa}}_n)^2. \tag{24.10}$$

This $Q^* \sim \chi^2_{K-1}(\lambda^*)$, where $\lambda^* = \sum_k n_k(\kappa_k - \hat{\bar{\kappa}}_n)^2$. The evidence in Q^* for heterogeneity of the κ_k's is clearly $T_{Q^*} = h_{K-1}(Q^*) \sim N(\tau_{Q^*}, 1)$, from Definition 24.1. Note that Q^* and Q do not usually measure the same type of heterogeneity. Further note that if all n_k's are equal, Q^* can be written $Q^* = \sum(T_k - \bar{T})^2$.

24.2 Simulation studies

For each of K studies one has n_k independent observations in the kth study, each with mean μ_k, variance σ_k^2. Let $\hat{\mu}_k$ be the sample mean, so $w_k^{-1} = \text{Var}[\hat{\mu}_k] = \sigma_k^2/n_k$. Let $\hat{w}_k = n_k/s_k^2$, where s_k^2 is the sample variance of the n_k observations in study k. In this section let $Q = \sum_k w_k(\hat{\mu}_k - \bar{\mu}_w)^2$ denote Cochran's formula in the idealized situation where the weights are known, $\hat{Q}$ the usual formula with estimated weights (24.1) and Q^* given by (24.10).

In this simulation experiment, the evidences for heterogeneity in each of Q, $\hat{Q}$ and Q^* were found for each of 40 000 replications of the parameter settings. Table 24.1 shows the result for $K = 3$ groups with equal sample sizes from standard normal distributions. Thus the hypothesis of homogeneous means holds. The computation of Q assumes known weights $w_k = n_k/\sigma_k^2 = n_k$, while for $\hat{Q}$ the weights are estimated by $\hat{w}_k = n_k/s_k^2$.

Column 2 of Table 24.1 shows the sample sizes, and column 3 the empirical size of the nominally level 0.05 test, an estimate of $P(\chi^2_2 \geq 5.99) = 0.05$. Columns 5 and

Table 24.1 Empirical study of T_Q, $T_{\hat{Q}}$ and T_{Q^*} based on 40 000 samples from $K = 3$ standard normal populations, with equal sample sizes.

	n_k	Size	λ_Q	$\bar{Q}$	s_Q	$\bar{T}_Q$	s_{T_Q}
Q	5	0.05	0	1.98	1.99	−0.52	0.99
	10	0.05	0	1.99	2.00	−0.52	0.99
	20	0.05	0	2.00	2.02	−0.52	1.00
	40	0.05	0	1.99	1.99	−0.52	0.99
	80	0.05	0	2.00	1.99	−0.51	1.00
$\hat{Q}$	5	0.12	0	2.85	3.95	−0.28	1.21
	10	0.08	0	2.31	2.60	−0.42	1.08
	20	0.07	0	2.15	2.30	−0.47	1.04
	40	0.06	0	2.06	2.11	−0.49	1.01
	80	0.05	0	2.03	2.04	−0.50	1.01
Q^*	5	0.10	0	2.54	3.21	−0.37	1.15
	10	0.07	0	2.15	2.36	−0.47	1.04
	20	0.06	0	2.07	2.17	−0.50	1.02
	40	0.05	0	2.02	2.06	−0.51	1.01
	80	0.05	0	2.01	2.01	−0.51	1.00

6 give the empirical mean and standard deviation of Q, estimates of $\mathrm{E}[Q] = 2$ and $\mathrm{SD}[Q] = 2$. Columns 7 and 8 give the empirical mean and standard deviation of the evidence T_Q. Note that s_{T_Q} is near 1, as expected, and the mean evidence is slightly negative, indicating a small positive evidence for the null hypothesis of homogeneity.

Next consider the results for $\hat{Q}$ in the same table. It is clear from consideration of the empirical size mean and standard deviation that this $\hat{Q}$, the one actually used in practice, has a distribution which is shifted to the right of its limiting distribution χ^2_2. Nevertheless, the evidence $T_{\hat{Q}}$ appears to reliably point to the null hypothesis.

The results for Q^* in the same table are worth comparing with those of Q and $\hat{Q}$. For although Q^* measures the heterogeneity of the transformed effects $\kappa_k = \mathcal{K}(\delta_k)$, where $\delta_k = (\mu_k - \mu_0)/\sigma_k$, when the σ_k's are equal the δ_k's (and κ_k's) differ if and only if the μ_k's differ. Thus Q^* indirectly measures the evidence for heterogeneity of the μ_k's. Note that the size, mean and standard deviation of Q^*, as well as the mean and standard deviation of the evidence in T_{Q^*} are closer to those of the ideal Q than the commonly used $\hat{Q}$.

In Table 24.2 are shown the corresponding results for the three test statistics under the alternative of heterogeneity: here the means are $(0, 0, 1)$. All results are computed as above, but now the estimated level is replaced by the estimated 'power' at level 0.05; of course when the level is not 0.05 (see Table 24.2 for the actual size) it is the estimated power at a level equal to the size.

Consider first the results for the ideal Q which assumes known weights. The estimated means and standard deviations in columns 5 and 6 are estimates of

Table 24.2 Empirical study of T_Q, $T_{\hat{Q}}$ and T_{Q^*} based on 40 000 samples from $K = 3$ normal distributions, with means (0,0,1) and variances (1,1,1), again with equal sample sizes. The 'power' of each Q-test is not necessarily at level 0.05, but rather at the estimated size of the test shown in Table 24.1.

	n_k	'Power'	λ_Q	$\bar{Q}$	s_Q	$\bar{T}_Q$	s_{T_Q}
Q	5	0.35	3.33	5.32	4.13	0.65	1.12
	10	0.63	6.67	8.66	5.53	1.43	1.09
	20	0.92	13.33	15.38	7.55	2.54	1.03
	40	1.00	26.67	28.57	10.46	4.05	1.01
	80	1.00	53.33	55.26	14.67	6.20	1.00
$\hat{Q}$	5	0.44	3.33	7.73	9.11	1.01	1.50
	10	0.65	6.67	10.13	7.86	1.62	1.28
	20	0.92	13.33	16.52	9.21	2.65	1.17
	40	1.00	26.67	29.56	12.00	4.12	1.12
	80	1.00	53.33	56.24	16.32	6.25	1.10
Q^*	5	0.34	2.89	5.73	5.73	0.66	1.28
	10	0.56	5.78	8.07	5.88	1.28	1.15
	20	0.88	11.56	13.74	7.42	2.29	1.06
	40	0.99	23.12	25.10	10.01	3.69	1.02
	80	1.00	46.25	48.25	13.82	5.70	1.01

$\mathrm{E}[Q] = 2 + \lambda_Q$ and $\mathrm{SD}[Q] = \sqrt{4 + 4\lambda_Q}$. For example, when sample sizes are all equal to 5, the theoretical $\mathrm{E}[Q] = 5.33$ and its estimate is 5.32; and the theoretical $\mathrm{SD}[Q] = 4.16$ and its estimate 4.13. If the transformation to evidence T_Q worked perfectly, its standard deviation would be 1, but in fact it is slightly larger, near 1.12. The estimated expected evidence for heterogeneity is 0.65, with standard error 1.12, which is very weak for these sample sizes.

Now if one uses $\hat{Q}$, one sees that the power and evidence are exaggerated over what would one expect from using the asymptotic distribution under alternatives, namely that of Q. This limiting noncentral chi-squared distribution does not describe the actual distribution of $\hat{Q}$.

Finally, for Q^* the noncentrality parameter λ_{Q^*} is smaller than that of λ_Q, but the noncentral chi-squared parameters $\nu = K - 1 = 2$ and λ_{Q^*} yield theoretical mean and standard deviation that are in good agreement with the estimated mean and standard deviation, especially for sample sizes at least 10 each. Thus the statistic T_{Q^*} can be relied upon as measure of evidence for heterogeneity of standardized effects, and in the case of equal variances, of the effects themselves.

The null distribution of Q^* is quite stable under changes of the parameters, especially compared to $\hat{Q}$. For example, if the null hypothesis of homogeneity still holds for $K = 7$ groups, and the smallest samples correspond to the populations with the

Table 24.3 Empirical study of T_Q, $T_{\hat{Q}}$ and T_{Q^*} based on 40 000 samples of respective sample sizes $n = (5, 5, 5, 10, 10, 10, 25)$ from $K = 7$ normal populations having means $\mu = (0, 0, 0, 0, 0, 0, 0)$ and standard deviations $\sigma = (2, 2, 2, 1, 0.5, 0.5, 0.5)$. These results are shown in rows 2, 7 and 12, which are labeled $\bar{n} = 10$. The subsequent rows show the results as the sample sizes are repeatedly doubled.

	$\bar{n}$	Size	λ_Q	$\bar{Q}$	s_Q	$\bar{T}_Q$	s_{T_Q}
Q	10	0.05	0	6.00	3.48	−0.28	0.96
	20	0.05	0	5.98	3.44	−0.28	0.95
	40	0.05	0	6.01	3.49	−0.28	0.96
	80	0.05	0	6.01	3.49	−0.28	0.96
	160	0.05	0	5.99	3.46	−0.28	0.96
$\hat{Q}$	10	0.21	0	9.44	10.64	0.31	1.46
	20	0.11	0	7.07	4.76	−0.05	1.09
	40	0.08	0	6.48	3.99	−0.17	1.02
	80	0.06	0	6.24	3.73	−0.23	0.99
	160	0.06	0	6.09	3.57	−0.26	0.97
Q^*	10	0.11	0	7.05	5.17	−0.09	1.15
	20	0.07	0	6.31	3.93	−0.22	1.01
	40	0.06	0	6.14	3.68	−0.25	0.98
	80	0.06	0	6.08	3.59	−0.27	0.97
	160	0.05	0	6.02	3.51	−0.28	0.96

largest variability, then the null distribution of $\hat{Q}$ is again shifted to the right much more than that of Q^* (see Table 24.3).

Other simulation studies were carried out, using $\hat{Q}$ and Q^* as defined above, but the data were not generated according to the normal model; rather they were generated from the symmetric but heavy tailed Student's t_3 model and the double exponential model; and also an asymmetric model composed of 80 % normal and 20 % from a standardized exponential model. In all these cases the null distribution of Q^* was closer to the nominal χ^2_{K-1} distribution than that of $\hat{Q}$ (see Kulinskaya and Staudte (2007) for details).

24.3 Cochran's Q: the random effects model

In the fixed effects model (FEM) of Section 24.1 it was assumed that there existed K independent $\hat{\mu}_k \sim N(\mu_k, w_k^{-1})$, where $w_k = \sigma_k^2/n_k$, either exactly, or in the limit as all $n_k \to \infty$. If this model is considered conditional on the μ_k's themselves being a random sample from the $N(\mu, \gamma^2)$ distribution, then the unconditional distribution of the $\hat{\mu}_k$'s is called the *random effects model* (REM). For the REM each $\hat{\mu}_k \sim N(\mu, w_k^{-1} + \gamma^2)$, where $\gamma^2 \geq 0$.

The parameter γ^2 is called the interstudy variance component, and under the null hypothesis $\gamma = 0$ the common distribution of the $\hat{\mu}_k$'s is the same as for the FEM under the null hypothesis of equal effects $\mu_k = \mu$. Thus the null distribution of Cochran's Q is the central χ^2_{K-1} under either model.

However, the alternative hypothesis $\gamma > 0$ under the REM is different from that of the FEM, which is that at least two of the μ_k's differ. These alternative hypotheses for both models describe 'heterogeneity', but for the FEM it is of fixed effects, while for the REM it is of random effects. Thus one cannot expect the evidence in Cochran's Q for heterogeneity to be the same for the two models.

Let $M_r = \sum_k w_k^r$ be the sum of rth powers of the weights, and define $a = M_1 - M_2/M_1$ and $b = M_2 - 2M_3/M_1 + (M_2/M_1)^2$, $c = b/a^2$ and $d = c(K-1) - 1$. All these constants are non-negative because the weights are assumed positive; and, when the weights are equal, $c = 1/(K-1)$ and $d = 0$. For the special case of $K = 2$, the constants $c = 1$ and $d = 0$ for any weights.

For the REM Biggerstaff and Tweedie (1997) obtain the moments

$$\mathrm{E}[Q] = K - 1 + a\,\gamma^2; \tag{24.11}$$
$$\mathrm{Var}[Q] = 2(K-1) + 4a\,\gamma^2 + 2b\,\gamma^4,$$

and approximate the distribution of Q by the gamma distribution with these moments. Here we want a *vst* for Q, so we write $\mathrm{Var}(Q) = 2d(K-1) - 4d\,(\mathrm{E}[Q]) + 2c\,(\mathrm{E}[Q])^2$. Thus $\mathrm{Var}[Q] = g(\mathrm{E}[Q])$, where $g(t) = a_0 + a_1 t + a_2 t^2$ and $a_0 = 2(K-1)d$, $a_1 = -4d$ and $a_2 = 2c$.

By the traditional method of Section 17.2 any indefinite integral $\int^x \{g(t)\}^{-1/2}\,\mathrm{d}t$ is a possible candidate for a vst . This requires a standard integral:

$$\int^x \{a_0 + a_1 t + a_2 t^2\}^{-1/2}\,\mathrm{d}t = \frac{1}{\sqrt{a_2}} \sinh^{-1}\left(\frac{2a_2 x + a_1}{\sqrt{4a_0 a_2 - a_1^2}}\right).$$

After substitution of the constants into this formula, we obtain

$$h(x) = \frac{1}{\sqrt{2c}} \sinh^{-1}\left(\frac{cx - d}{\sqrt{d}}\right).$$

This $h(x)$ has been chosen to stabilize the variance of Q at 1. To obtain a potential measure of evidence, one must also subtract off the mean $h(\mathrm{E}[Q])$ at the null hypothesis $\gamma = 0$; that is, when $\mathrm{E}[Q] = K - 1$. This leads to $T'_Q = h(Q) - h(K-1)$ as evidence for the alternative hypothesis $\gamma > 0$. Recalling the definition $\sinh^{-1}(y) = \ln\big(y + \sqrt{1 + y^2}\,\big)$, this evidence can be rewritten in terms of the log function.

Definition 24.2 *Assuming the random effects model as defined above, the evidence T'_Q in Cochran's Q for the alternative $\gamma > 0$ is defined by*

$$T'_Q = \frac{1}{\sqrt{2c}}\left\{\ln\left(\frac{cQ - d + \sqrt{(cQ-d)^2 + d}}{1 + \sqrt{1+d}}\right)\right\}. \tag{24.12}$$

For the case of equal weights $d = 0, c = 1/(K-1)$ *and* $T'_Q = (1/\sqrt{2c})\ln\big(cQ\big)$.

Substituting the expected value of Q given by (24.11) into (24.12) leads to the first-order approximation:

$$\mathrm{E}[T'_Q] \doteq \frac{1}{\sqrt{2c}}\left\{\ln\left(\frac{1 + ac\,\gamma^2 + \sqrt{d + (1 + ac\,\gamma^2)^2}}{1 + \sqrt{d+1}}\right)\right\}. \tag{24.13}$$

For the case of equal weights $d = 0, c = 1/(K-1)$ *and* $\mathrm{E}[T'_Q] \doteq (1/\sqrt{2c})\ln(1 + ac\,\gamma^2)$.

Remarks

1. A limited number of simulation studies were carried out using the REM to determine whether T'_Q has a stable variance for $0 \le \gamma \le 1$, a range including

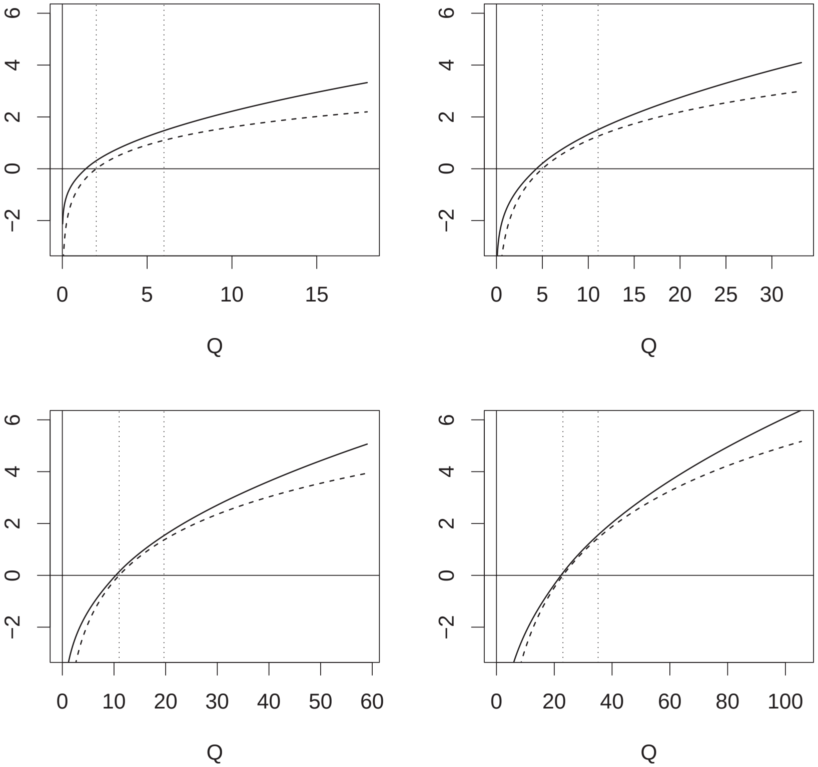


Figure 24.1 All plots compare the evidence in Q assuming the FEM (solid line) with that assuming the REM (dashed line), for $0 < Q \le 3\chi^2_{K-1,0.95}$. The top left-hand plot is based on $K = 3$ samples of size 10; the top right-hand plot has $K = 6$ samples of size 10; and K continues to double in the bottom left and right plots. The vertical dotted lines indicate the df K-1 and the critical point $\chi^2_{K,0.95}$ of the traditional test for heterogeneity.

all applications we have encountered. For equal weights $w_k = n_k = N/K$, the standard deviation of T'_Q was near 1.5 for $K = 2$, 1.3 for $K = 3$, 1.15 for $K = 5$ and 1.07 for $K = 10$, uniformly in γ over the range of interest. For unequal sample sizes the standard deviation varied slightly more about these values. Assuming equal sample sizes, the graph of the empirical mean evidence for heterogeneity versus γ had the same shape as the expected evidence (24.13). It was biased downwards, by the fixed amount, $1/\sqrt{K}$, and when this was added to T'_Q the bias almost disappeared. These results depend much more on the value of K than on the total sample size $N = \sum n_k$, assuming only all $n_k \geq 10$.

2. Assuming known weights it is of interest to compare the evidence for heterogeneity T_Q of Definition 24.1 for the FEM to the evidence in (24.12) derived

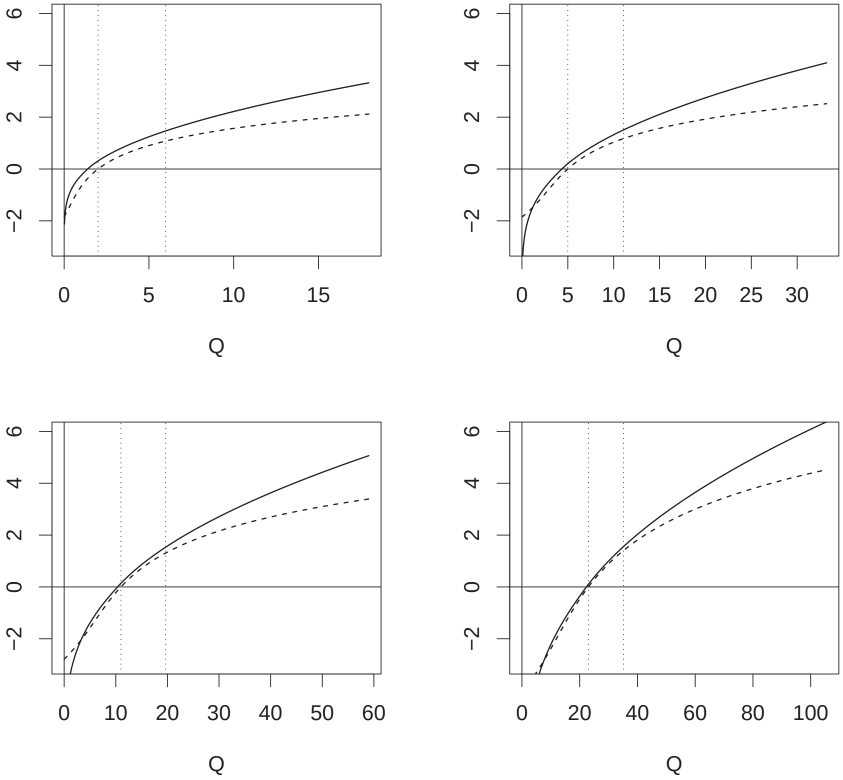


Figure 24.2 These plots are based on the same values of K as in Figure 24.1, but now the sample sizes are very unbalanced. For the top left-hand plot, they are (10, 10, 100); for the top right-hand plot (10, 10, 100, 10, 10, 100); and the pattern of doubling the number of studies with sample sizes (10,10,100) continues. The discrepancy in evidence for heterogeneity between the FEM and REM appears to be greater for this unbalanced case.

for the REM, because the same Q is often used to test for heterogeneity in both models. Some plots making direct comparisons of these measures of evidence are shown in Figures 24.1 and 24.2. Consider the upper left-hand plot of Figure 24.1. It shows for $K = 3$ samples of equal size $n_k = 10$ the graph of T_Q versus Q as a solid line compared to the graph of T'_Q versus Q as a dashed line. The vertical dotted lines indicate the df $K - 1$ and the critical point $\chi^2_{K-1,0.95}$ of the traditional test for heterogeneity. Note that both graphs indicate similar evidence for their respective alternative hypotheses for Q up to this critical point, but that for larger Q the evidence for heterogeneity is lower in the REM.

3. The plots of Figures 24.1 and 24.2 do not depend on the size of $w_k = n_k$; rather it is the configuration of weights that matters, as well as the value of K.

4. The derivation leading to the above definition of T'_Q depends on the assumption of *known* weights, and its validity can be compromised by substitution of estimates for them. In applications one applies these evidences for heterogeneity to transformed effects, which are supposed to be approximately normally distributed, with known standard variances $1/n_k$ under the FEM. But this will typically only be the case for all $n_k \geq 10$, and the larger the n_k's in the individual studies, the better.

5. One could define a new version of Cochran's Q using weights $\hat{w}_k^{-1} + \hat{\gamma}^2$, but γ^2 is not easy to estimate, especially when it is small (see Chapter 25).

24.4 Summary

In this chapter we studied Cochran's Q under the ideal situation where the weights are known and when they are estimated. We showed that the study sample sizes n_k must be quite large before estimated weights can be safely substituted in Q. That is why we advocate thinking in terms of transformed standardized effects, whose distributions are designed to be approximately normal with variances equal to the reciprocals of the sample sizes. Then the weights on the transformed space are 'known' to be $w_k = n_k$, and one uses the special case of Cochran's Q that is denoted Q^*. We also showed that evidence for heterogeneity in Cochran's Q depends on which model is used: for the random effects model T'_Q increases in Q as $\ln(Q)$; while for the fixed effects model T_Q increases at the faster rate $\sqrt{Q}$.

25

Combining evidence from K studies

25.1 Background and preliminary steps

In this book we have shown that a *vst* $T = h(S)$ of a test statistic for a positive effect $\mu > 0$ often can be chosen so that the T lies on the unit normal calibration scale; that is, to a useful approximation $T \sim N(\tau, 1)$. This construction allows one to easily interpret the evidence in the test statistic S for $\mu > 0$, because T is an estimator of its expected evidence τ with known standard normal error. Frequently the test statistic S and the *vst* h can be chosen so that $T \sim N(\sqrt{n}\,\kappa, 1)$, where n is the sample size, and $\kappa = \mathcal{K}(\delta)$ is the Key Inferential Function applied to the standardized effect δ. This construction allows one not only to find confidence intervals for κ, but also, by back-transformation, for δ.

Now suppose there are K independent studies with data which, it is decided, can be interpreted using the same model with unknown parameters. The parameter of interest, the effect, may change from study to study, so it is denoted μ_k for the kth study. Often there are other parameters which may vary. For example if the normal model is adopted, both μ_k and σ_k may differ with k, and if the Student t-statistic S_k is used in the kth study, we found that a *vst* led to evidence whose mean grew with a monotonic function of the standardized effect $\delta_k = (\mu_k - \mu_0)/\sigma_k$.

Let $\kappa_k = \mathcal{K}(\delta_k)$ denote the kth transformed effect. In Chapter 24 we applied Cochran's Q to the $\hat{\kappa}_k = \mathcal{K}(\hat{\delta}_k)$, $k = 1, \ldots, K$ to obtain the evidence T_{Q^*} for heterogeneity of the κ_k's directly, and the δ_k's indirectly. On the basis of this evidence one then has to make a decision how to proceed. If there is little or no evidence (say $T_{Q^*} < 1.645$) for heterogeneity of fixed effects, then one might assume equal

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

standardized effects $\delta_k = \delta$ for all k, and proceed to combine evidence for the alternative $\delta > 0$ and find confidence intervals for δ as described in Section 25.2.1.

Now assume T_{Q^*} is large enough to raise doubts about the above simple model. We describe three ways to proceed, depending on the assumptions one is willing to make and the data available for analysis. First, if one wants to estimate a fixed and representative standardized effect δ for the K studies, one can proceed as in Section 25.2.2. Second, if one assumes the K studies in hand are a random sample from a larger population of studies, present or future, and wants to draw inferences about a representative δ for this larger population, then one can proceed as in Section 25.3. Third, if one suspects that a covariate can explain the differences in the δ_k, then one can employ meta-regression as explained is Chapter 14.

25.2 Fixed standardized effects

25.2.1 Fixed, and equal, standardized effects

Given K independent studies, with evidence T_k in the kth study for $\delta > 0$, for $k = 1, \dots, K$. Then, at least approximately, each $T_k \sim N(\tau_k, 1)$ where $\tau_k = \sqrt{n_k}\,\kappa$, $\kappa = \mathcal{K}(\delta)$ and $\mathcal{K}$ is the Key Inferential Function for the assumed model. The combined evidence for $\delta > 0$ should continue to be on the evidence scale: that is, it should continue to be approximately normally distributed with variance 1, and mean growing with δ. By the method of Lagrange multipliers one can show that amongst all linear combinations $\sum_k v_k T_k$, all $v_k > 0$, satisfying $\mathrm{Var}[\sum_k v_k T_k] = \sum v_k^2 = 1$, choosing v_k proportional to $\sqrt{n_k}$ maximizes the expected evidence $\mathrm{E}[\sum_k v_k T_k] = \sum_k v_k \tau_k \doteq \left(\sum_k v_k \sqrt{n_k}\right)\kappa$. Therefore we choose this combination.

Definition 25.1 *Define the* combined evidence *for $\delta > 0$ in the K studies by*

$$T_{1:K} = \frac{\sqrt{n_1}\,T_1 + \cdots + \sqrt{n_K}\,T_K}{\sqrt{n_1 + \cdots + n_K}}. \tag{25.1}$$

As usual, when $T_{1:K}$ is negative, its magnitude $|T_{1:K}|$ is interpreted as positive evidence for $\delta < 0$.

Now $\mathrm{E}[T_{1:K}] \doteq \sqrt{N}\,\kappa$, where $N = \sum_k n_k$. So a $100(1-\alpha)$ % confidence interval for κ is given by $(T_{1:K} \pm z_{1-\alpha/2})/\sqrt{N}$. An interval of the same confidence for $\delta = \mathcal{K}^{-1}(\kappa)$ is obtained by applying $\mathcal{K}^{-1}$ to the endpoints.

If for each k the evidence T_k is of the form $T_k = \sqrt{n_k}\,\mathcal{K}(\hat{\delta}_k)$ where $\hat{\delta}_k$ is an estimator of δ in the kth study, then the above interval can be reexpressed in terms of the $\hat{\kappa}_k = \mathcal{K}(\hat{\delta}_k)$'s. For then $T_{1:K}/\sqrt{N}$ is the weighted combination of the $\hat{\kappa}_k$'s, namely $T_{1:K}/\sqrt{N} = \hat{\kappa} = \sum_k n_k \hat{\kappa}_k / N \sim N(\kappa, 1/N)$.

The coverage of these intervals should be better than that of the coverage of intervals in individual studies, because weighted averaging of the T_k's should result in a distribution closer to normality.

25.2.2 Fixed, but unequal, standardized effects

In this section the δ_k's are not assumed equal, but a representative δ for these K studies is desired. To obtain one, we first define a weighted transformed effect, and then apply $\mathcal{K}^{-1}$ to it.

Definition 25.2 *Given $\kappa_k = \mathcal{K}(\delta_k)$, $k = 1, \ldots, K$, define a representative κ by*

$$\kappa = \sum_k n_k \kappa_k / N. \tag{25.2}$$

This κ gives weight proportional to the sample sizes involved in the K studies. Other weights may be more appropriate, depending on circumstances. Define the representative δ for the K studies by $\delta = \mathcal{K}^{-1}(\kappa)$, where $\mathcal{K}$ is the appropriate Key for the K studies.

The evidence for $\delta > 0$ is defined the same way it was for equal effects (Equation (25.1)).

To find an interval estimate for δ we first find one for κ. Starting with $\hat{\kappa} = \sum_k n_k \hat{\kappa}_k / N$, it is easy to see that $\hat{\kappa}$ is unbiased for κ, and further $\hat{\kappa} \sim N(\kappa, 1/N)$. As in the previous section of equal standardized effects, a nominal $100(1-\alpha)\,\%$ confidence interval for κ has endpoints $\hat{\kappa} \pm z_{1-\alpha/2}/\sqrt{N}$, and the same confidence can be had in the interval for δ obtained by applying $\mathcal{K}^{-1}$ to the endpoints of this interval.

The reader will have noticed that the estimation methodology is exactly the same as for fixed equal effects, but the parameter of interest δ now has a different meaning. While before δ was assumed fixed for all studies, now the δ_k's are allowed to vary, and δ is the standardized effect that transforms into the weighted average of the κ_k's. So the interpretation of δ is quite different.

25.2.3 Nuisance parameters

We have found that in some contexts the Key $\mathcal{K}$ depends not only on a standardized effect δ of interest but also on a nuisance parameter ξ; thus $\mathcal{K} = \mathcal{K}(\delta, \xi)$. For example with the two-sample t-test the Key depended on both the standardized difference of means δ and $\xi^{-1} = \nu/N$, the ratio of Welch's degrees of freedom and the total sample size (see Chapter 21 for details). In such cases one has not only δ_k's from K studies to combine, but also the ξ_k's. It appears that each problem may require an *ad hoc* solution, but if $\mathcal{K}(\delta, \xi)$ is monotonic in ξ, a weighted average of the ξ_k's, with weights proportional to the sample sizes, seems to be a useful prescription. Once a representative ξ is defined, the same combination of $\hat{\xi}_k$'s leads to an estimator $\hat{\xi}$ of ξ, and one can continue as follows.

For simplicity assume that $\mathcal{K}(\delta, \xi)$ is not only strictly monotonic in each argument, but jointly continuous in both arguments. Let $\kappa_k = \mathcal{K}(\delta_k, \xi_k)$ for all k, and let $\hat{\kappa}_k = T_k/\sqrt{n_k}$. This $\hat{\kappa}_k \sim N(\kappa_k, 1/n_k)$. For $N = \sum_k n_k$, define κ by (25.2). Then define an overall standardized effect δ in terms of the representative ξ, κ as the solution of

$$\mathcal{K}(\delta, \xi) = \kappa. \tag{25.3}$$

This δ exists because $\mathcal{K}(\delta, \xi)$ is monotonically increasing in its first argument, by definition. A nominal $100(1-\alpha)\,\%$ confidence interval for κ has endpoints $\hat{\kappa} \pm z_{1-\alpha/2}/\sqrt{N}$, and the same confidence can be found in an interval for δ obtained by fixing ξ at $\hat{\xi}$ and solving (25.3) for the endpoints of this interval. This procedure should be checked by simulations, as has been done for the two-sample normal model in Chapter 21 and in Kulinskaya and Staudte (2007).

Another attractive approach is to acknowledge different nuisance parameters and condition on their values. Then Equation (25.3) is changed to

$$N^{-1} \sum n_k \mathcal{K}(\delta, \xi_k) = \kappa. \tag{25.4}$$

The solution δ exists because the sum of the key functions is still monotonically increasing in δ, and the confidence interval for δ is obtained as above after fixing ξ_k at $\hat{\xi}_k$. This procedure has not yet been tried in applications.

25.3 Random transformed effects

25.3.1 The random transformed effects model

In traditional meta analysis interstudy variability is often modeled by assuming the effects themselves are a random sample from a normal model with a positive variance. We follow this example, but introduce the interstudy normal model on the space of transformed standardized effects, because on this space we have estimators resulting from variance stabilization, and we know they are approximately normal with known variances.

Given a transformation $\delta \to \mathcal{K}(\delta)$ that is monotonically increasing and continuous, define for each k the transformed effect $\kappa_k = \mathcal{K}(\delta_k)$. In the previous sections, it was assumed the κ_k's were fixed, but now it is assumed $\kappa_1, \ldots, \kappa_K$ are a sample from the $N(\kappa, \gamma^2)$ model with unknown mean κ and variance γ^2.

Conditional on the observed values of $\kappa_1, \ldots, \kappa_K$ it is assumed further that there exists estimators $\hat{\delta}_k$, $k = 1, \ldots, K$, for which $\hat{\kappa}_k = \mathcal{K}(\hat{\delta}_k)$ has a conditional distribution, given κ_k, which is $N(\mathcal{K}(\delta_k), 1/n_k)$. (When $\mathcal{K}$ is the $\sinh^{-1}$ transformation applied to Student t-statistics this amounts to assuming the δ_k's are independent, and from rescaled Student t-distributions.)

To obtain the unconditional properties of the $\hat{\kappa}_k$'s one must average over the distribution $N(\kappa, \gamma^2)$. By using the conditioning formulas for expectations and variances, one finds:

$$\begin{aligned} \mathrm{E}[\hat{\kappa}_k] &= \kappa \\ \mathrm{Var}[\hat{\kappa}_k] &= \frac{1}{n_k} + \gamma^2. \end{aligned} \tag{25.5}$$

These assumptions define the *random transformed effects model*. The first goal is to find evidence for $\kappa > 0$, and hence $\delta = \mathcal{K}^{-1}(\kappa) > 0$, where $\mathcal{K}$ is the common Key for the K studies. The second goal is to find a confidence interval for κ, and by back-transformation δ. To achieve these goals, one apparently needs an estimator of γ^2, but this is not the case, as we will see.

Before going any further, however, it is suggested that one find the evidence for $\gamma > 0$ using the evidence for heterogeneity T'_{Q^*} in the REM defined by (24.12) applied to $Q^* = \sum_k n_k(\hat{\kappa}_k - \hat{\kappa})^2$. For, even if T_{Q^*}, the evidence for heterogeneity of fixed effects is large, it does not mean the evidence for heterogeneity of random effects is large enough to worry about. Assuming there is weak or stronger evidence for $\gamma > 0$, one can then proceed as follows.

Let $\bar{\kappa} = (\sum_k \hat{\kappa}_k)/K$ and $s_\kappa^2 = \sum_k(\hat{\kappa}_k - \bar{\kappa})^2/(K-1)$ denote the sample mean and variance of the $\hat{\kappa}_k$'s. Then one can show that $\mathrm{E}[s_\kappa^2] = \gamma^2 + (1/K)\sum_k(1/n_k)$, which, together with the fact that $\gamma^2 \geq 0$, leads to the estimator of γ^2 :

$$\hat{\gamma}^2 = \max\left\{0,\ s_\kappa^2 - \frac{1}{K}\sum_k \frac{1}{n_k}\right\}. \tag{25.6}$$

In this context of known weights equal to the sample sizes, the DerSimonian and Laird (1986) estimator of γ^2 reduces to

$$\hat{\gamma}^2_{DL} = \max\left\{0,\ \frac{Q^* - (K-1)}{N - \sum_k n_k^2/N}\right\}, \tag{25.7}$$

where $N = \sum_k n_k$, $\hat{\kappa} = (\sum_k n_k\hat{\kappa}_k)/N$, and $Q^* = \sum_k n_k(\hat{\kappa}_k - \hat{\kappa})^2$. It is readily seen that when the weights n_k are equal, the above two estimators are identical.

25.3.2 Evidence for a positive effect

For the random transformed effects model just defined, the evidence $T_{1:K}$ of (25.1) is a linear combination of normally distributed variables $T_k = \sqrt{n_k}\,\hat{\kappa}_k$, and hence normally distributed, but its variance now depends on the unknown γ^2. And, substituting one of the estimators $\hat{\gamma}$ or $\hat{\gamma}_{DL}$ for γ^2 leads to $T_{1:K}$ with an unknown distribution. Therefore we proceed differently.

Note that $\bar{\kappa}$ is an unbiased estimator of κ with variance

$$\mathrm{Var}[\bar{\kappa}] = \frac{1}{K^2}\sum_k\left\{\frac{1}{n_k} + \gamma^2\right\} = \frac{\mathrm{E}[s_\kappa^2]}{K}\ , \tag{25.8}$$

using the fact that $\mathrm{E}[s_\kappa^2] = \gamma^2 + (1/K)\sum_k(1/n_k)$. This suggests the Studentized sample mean $S_{K-1} = \sqrt{K}\,(\bar{\kappa} - 0)/s_\kappa$ as a possible basis for measuring evidence for $\kappa > 0$, but it has an unknown distribution.

However, it is clear from (25.5) that the $\hat{\kappa}_k$'s will have constant variance whenever all $n_k = n$, say, and then $S_{K-1} \sim t_{K-1}(\lambda)$, the noncentral Student t-distribution with $K-1$ degrees of freedom and noncentrality parameter

$$\lambda = \sqrt{K}\,\frac{(\kappa - 0)}{\sqrt{1/n + \gamma^2}}. \tag{25.9}$$

The same result will hold to a good approximation if every $1/n_k$ is small relative to γ^2, because then the $\hat{\kappa}_k$'s are a random sample from (almost) the same normal population. Of course this speculation needs to be checked by simulations.

Let s_{1/n_k} be the standard deviation of the reciprocal sample sizes. For all $K \geq 2$, $\gamma > 2s_{1/n_k}$, the simulations described in Section 25.3.5 suggest that this is so; that is, the variances $1/n_k + \gamma^2$ of the $\hat{\kappa}_k$'s are sufficiently close to each other so that for all practical purposes one can proceed as though they were equal in the REM.

The condition $\gamma > 2s_{1/n_k}$ is likely to be met in practice if all $n_k \geq 10$, because then this is true for $\gamma > 0.05$. One can estimate γ using (25.6) or (25.7), for example, but caution is in order because these estimators are biased upwards for small γ. There are other estimators of γ available, including a MLE by Biggerstaff and Tweedie (1997), but to our knowledge its performance has not yet been checked with simulations.

Definition 25.3 *Assume the random transformed effects model for K studies, and assume that all study sample sizes are at least 10 and* $\gamma > 2s_{1/n_k}$. *The evidence for* $\kappa > 0$ *and hence* $\delta > 0$ *is given by applying the vst (20.4) for the noncentral t-distribution to the statistic* S_{K-1}:

$$T^*_{1:K} = \sqrt{2K}\ \sinh^{-1}\left(\frac{S_{K-1}}{\sqrt{2K}}\right) = \sqrt{2K}\ \sinh^{-1}\left(\frac{\bar{\kappa}}{\sqrt{2}\, s_\kappa}\right) . \qquad (25.10)$$

25.3.3 Confidence intervals for κ and δ: K small

Given the rationale for evidence for a positive effect in the REM in the previous section, it is now tempting to employ the Student t-interval with $c = t_{K-1,1-\alpha/2}$ to capture κ:

$$[L, U] = \left[\bar{\kappa} - c\,\frac{s_\kappa}{\sqrt{K}},\ \bar{\kappa} + c\,\frac{s_\kappa}{\sqrt{K}}\right]. \qquad (25.11)$$

Simulation studies described in Section 25.3.5 indicate that these t-intervals for κ have very accurate coverage for every $K > 1$ and all $\gamma > s_{1/n_k}$.

The small sample confidence interval for δ is then given by $[\mathcal{K}^{-1}(L), \mathcal{K}^{-1}(U)]$, where $\mathcal{K}$ is the common Key for the K studies.

25.3.4 Confidence intervals for κ and δ: K large

In the previous section we estimated κ using equal weights on each $\hat{\kappa}_k$, but one may want more weight on $\hat{\kappa}_k$'s which are based on larger sample sizes. Let $\hat{\kappa}_v = \sum_k v_k\hat{\kappa}_k / \sum_j v_j$ be an estimator of κ with known positive weights v_k. Then $\hat{\kappa}_v \sim N(\kappa, \sigma_v^2)$, where

$$\sigma_v^2 = \frac{1}{\{\sum_j v_j\}^2}\left[\sum_k v_k^2 \left\{\frac{1}{n_k} + \gamma^2\right\}\right]. \qquad (25.12)$$

For inverse variance weights $v_k = n_k$, and $N = \sum_k n_k$ this simplifies to $\sigma_v^2 = (1/N + \gamma^2 \sum_k n_k^2/N^2)$. Letting $c = z_{1-\alpha}$, a large-sample $100(1-\alpha)$ % confidence interval for κ is given by

$$[L_2, U_2] = [\hat{\kappa}_v - c\,\hat{\sigma}_v,\ \hat{\kappa}_v + c\,\hat{\sigma}_v], \tag{25.13}$$

where $\hat{\sigma}_v$ is obtained from (25.12) after estimating γ^2 using (25.6) or (25.7). These last estimates require K to be very large, in order to achieve the nominal 95 % coverage, as demonstrated in the next section.

The large sample confidence interval for δ is then given by $[\mathcal{K}^{-1}(L_2), \mathcal{K}^{-1}(U_2)]$, where $\mathcal{K}$ is the common Key for the K studies.

25.3.5 Simulation studies

In order to evaluate the performance of the confidence intervals described in the previous two sections, a variety of values of K and sample sizes $n_1, \ldots, n_K$ were chosen. For each of these choices, 40 000 simulated samples $\hat{\kappa}_1, \ldots, \hat{\kappa}_K$ were generated with $\kappa_k \sim N(\kappa, 1/n_k + \gamma^2)$, where the target κ was held fixed, and γ set initially to 0. The three intervals initially compared were:

1. the 95 % Student t-interval defined by (25.11);
2. the large-sample 95 % interval defined by (25.13), with $c = z_{0.975}$ and γ estimated by $\hat{\gamma}$ of (25.6); and
3. the large-sample 95 % interval defined by (25.13), again with $c = z_{0.975}$ and γ estimated by $\hat{\gamma}_{DL}$ of (25.7).

This sampling procedure was repeated for 30 more selected values of γ in the unit interval. This region includes all estimated values of γ we have seen in applications; and, in any case, simulations for γ ranging from 1 to 20 yielded no changes from those at $\gamma = 1$. The resulting empirical coverage probabilities for the three intervals were plotted as functions of γ.

- It immediately became apparent that the asymptotic intervals, points (2) and (3) above, had coverage less than 95 %, sometimes by a large margin, for a wide range of values of γ, unless K was at least 40. For example, when $K = 30$ and all $n_k = 10$, these intervals have coverage near 96 % for very small γ, but this drops to 94 % for $0.2 \leq \gamma \leq 1$. Increasing all n_k's does not improve the coverage; it is the value of K that must be increased. Thus the asymptotics do not 'kick in' early enough for these intervals to be of practical value. Their performance was greatly improved by replacing $c = z_{0.975}$ by $c = t_{K-1,0.975}$, so hereafter we make this change.
- For small K the coverages of the t-intervals were overly conservative for very small γ but performed extremely well otherwise. The example with $n_1 = n_2 = 10$, $n_3 = 50$ is displayed in the top plot of Figure 25.1.

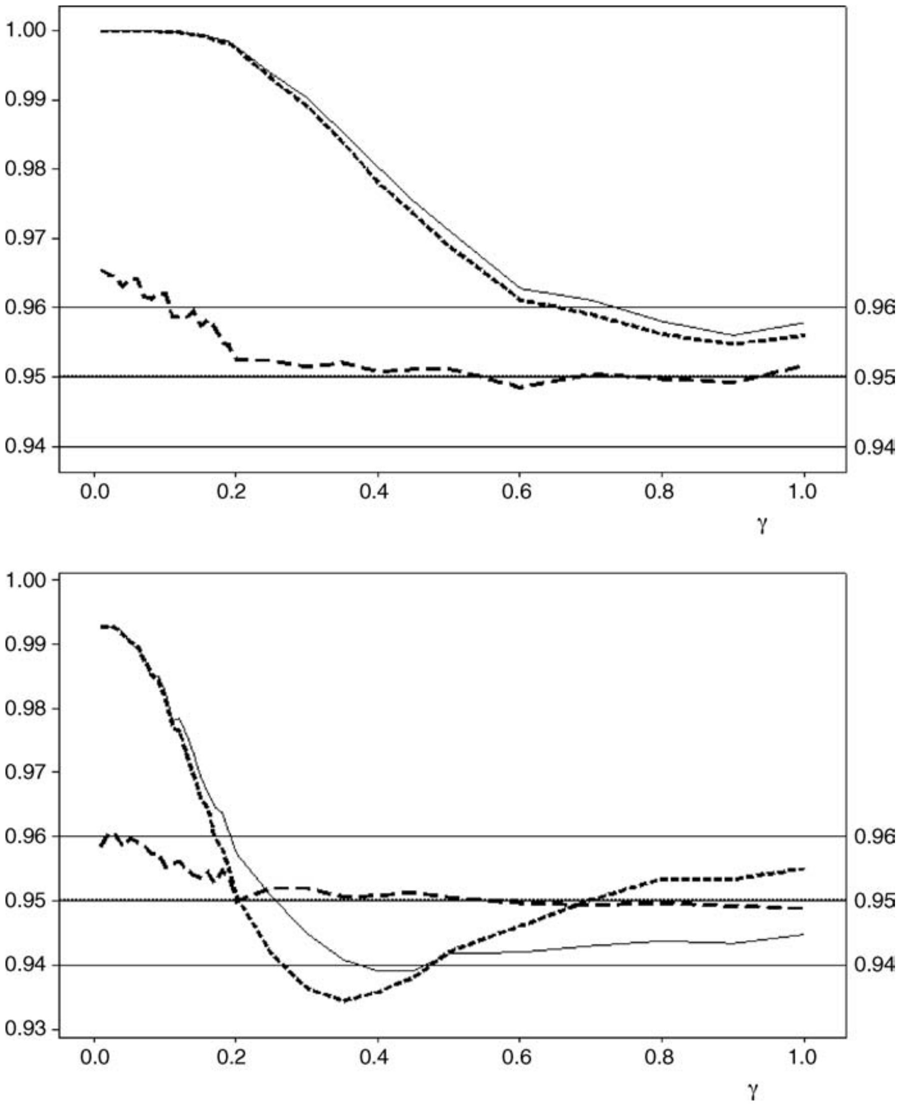


Figure 25.1 The top plot is for $K = 3$ studies of sizes 10, 10 and 50. The dashed line gives the empirical coverage of the t-interval defined by (25.11) as a function of γ. The other graphs depict coverage of intervals defined by (25.13), with the dotted line corresponding to $\hat{\gamma}$ of (25.6) and the solid line to $\hat{\gamma}_{DL}$ of (25.7). In both cases $c = t_{2,0.975}$ as in (25.11). In the bottom plot are shown the graphs of the coverage probabilities of these three intervals for $K = 6$ studies of sizes 10, 10, 10, 50, 50 and 50.

- The lower plot of Figure 25.1 shows the results for $K = 6$ studies with $n_1 = n_2 = n_3 = 10$ and $n_4 = n_5 = n_6 = 50$. This time the $\hat{\gamma}_{DL}$-based interval fares much better than that based on $\hat{\gamma}$, but both are soundly beaten by the Student t-interval.
- The upper plot of Figure 25.2 compares the coverages of these intervals for the sample sizes of the 11 studies of the recurrent urinary tract infections data described in Section 19.5.
- The lower plot of Figure 25.2 assumes 30 studies having sample size 10 and 10 having sample size 50. Even for this number of studies, the t-interval has the best overall performance.
- The average lengths of the intervals were also found, and the t-intervals were shorter than the other two intervals, which were similar in length.

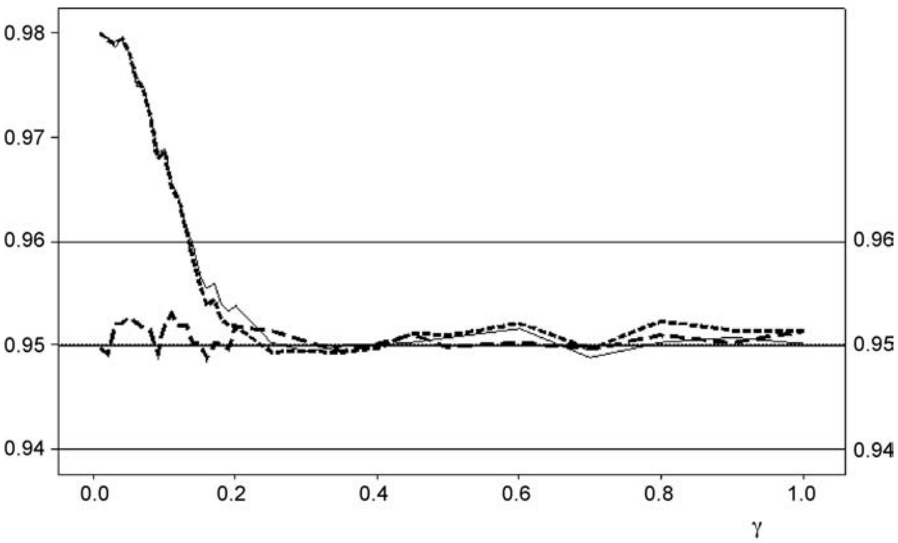
In summary, for the ideal situation where the transformed effects $\hat{\kappa}_k$ are exactly normally distributed with variances $1/n_k + \gamma^2$, the Student t-intervals for κ are preferred. The second best performer was the large-sample interval centered on a weighted estimator $\hat{\kappa}$ and using the DerSimonian and Laird (1986) estimator of γ. However, it needed to be modified, replacing $z_{0.975}$ by $t_{K-1,0.975}$, for it to be competitive unless K is at least 40. Estimation of γ is not necessary to carry out inference regarding κ for the REM, as we have seen. For those readers who want a confidence interval for the parameter γ, we suggest Biggerstaff and Tweedie (1997).

In practice the transformed effects $\hat{\kappa}_k$ will only be approximately normal with standard deviations approximately $1/n_k + \gamma^2$ under the random (transformed) effects model, so the above results must be treated with caution. Sample sizes in individual studies must be large enough for variance stabilization techniques to work, and how large they must be depends on the model and (unknown) values of the parameters. This warning also applies to other meta-analytic techniques that use *estimated* weights, especially ones that advocate normal approximations for K only moderately large.

25.4 Example: drop in systolic blood pressure

We return to the example of two-sample comparisons studied in Section 21.1, with original data in Table 21.1 and results for individual studies summarized in Table 21.3. Recall that N_k is the total sample size in the kth study, T_k is the evidence for a positive effect and $\hat{\kappa}_k = T_k/\sqrt{N_k}$ is the estimated transformed standardized effect. These last two results are shown in Table 25.1 to three decimal places. In this example Cochran's Q applied to the transformed effects yields $Q^* = \sum n_k(\hat{\kappa}_k - \hat{\kappa})^2 = 14.035$ which exceeds $\chi^2_{6,0.95} = 12.6$, so this traditional test would reject the assumption of equal transformed standardized effects; that is, the assumption of equal κ_k's is rejected at level 0.05 by this test. This is custom, but still arbitrary.

Let $m_6 = \chi^2_{6,0.5} = 5.34812$ be the median of the χ^2_6 distribution. The evidence for heterogeneity T_{Q^*} is found by applying the *vst* (22.1) to Q^*; it yields $T_{Q^*} = \sqrt{Q^* - m_6/2} - \sqrt{m_6/2} = 1.7$, which is only weak evidence for heterogeneity of the fixed standardized effects. Without further information it could reasonably be



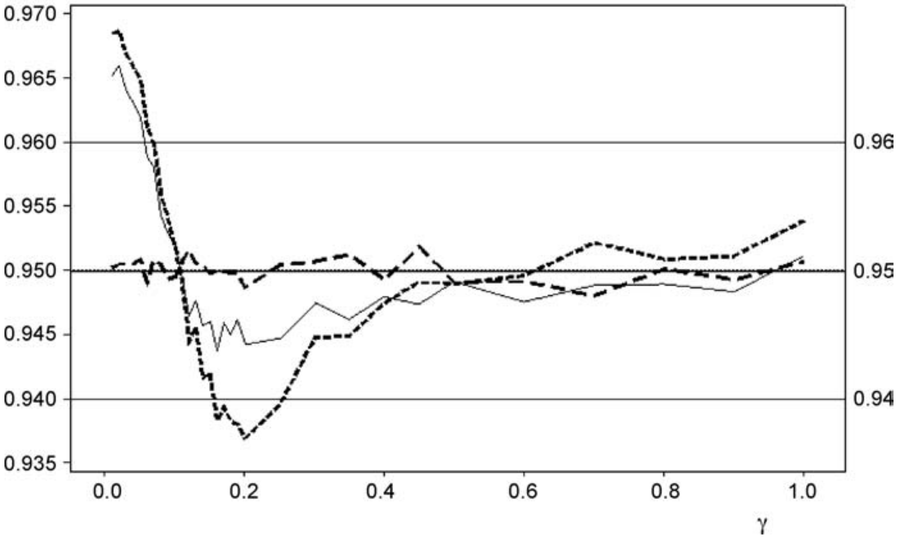


Figure 25.2 Continuing with the same intervals as in Figure 25.1, the top plot shows empirical coverage for $K = 11$ studies with sample sizes 45, 40, 28, 41, 24, 35, 19, 50, 43, 20 and 27. Only the t-interval coverage is close to 0.95 for all γ. Similar results are obtained in the bottom plot for $K = 40$ studies, 30 having sample size 10 and 10 having sample size 50. The average lengths of these intervals were also computed and are substantially smaller for the t-intervals compared to the others, especially for small K.

Table 25.1 For each of seven studies are shown the total sample sizes N_k, the Welch df $\hat{\nu}_k$, evidence for positive effect T_k and corresponding transformed effects $\hat{\kappa}_k$.

k	N_k	$\hat{\nu}_k$	T_k	$\hat{\kappa}_k$
1	51	48.30	−1.263	−0.177
2	38	35.58	2.158	0.350
3	130	127.62	2.404	0.211
4	19	17.00	0.997	0.229
5	49	44.24	−1.285	−0.184
6	10	7.26	1.608	0.509
7	33	27.91	1.117	0.195

assumed that the standardized effects are fixed and unequal, or that they are a random sample from a population of standardized effects.

The choice of model (fixed or random standardized effects) should be determined on the basis of whether one wants to draw inferences regarding these seven studies only, or rather one wants to draw inferences for a larger population of studies for which these seven represent a genuine random sample. We will do the computations for each model for illustrative purposes.

25.4.1 Inference for the fixed effects model

Whether one assumes all $\kappa_k = \kappa$ or the κ_k's are different, and the representative $\kappa = \sum_k N_k \kappa_k / N$, where $N = \sum_k N_k$, the inferential methods are the same. Using (25.1) the evidence for $\kappa > 0$ in these $N = 7$ studies is $T_{1:7} = 2.12$, which is weak. That is, there is only weak combined evidence for the conclusion that dieting leads to a drop in systolic blood pressure. This is not surprising, given that two of the seven studies showed the opposite result. This evidence is readily converted into a 95 % interval $(T_{1:7} \pm z_{0.975})/\sqrt{7}$ for κ, namely [0.009, 0.225].

A 95 % confidence interval for $\delta = \mathcal{K}^{-1}(\kappa)$ requires the inverse of the Key for the Welch t-test. Recall from Chapter 21 the Key $\mathcal{K}(\delta)$ is for each value of the nuisance parameter ξ given by

$$\mathcal{K}_\xi(\delta) = \sqrt{\frac{2}{\xi}}\,\sinh^{-1}\left(\frac{\sqrt{\xi}\,\delta}{\sqrt{2}}\right).$$

In the kth study $\xi_k = N_k/\hat{\nu}_k$, the ratio of the total sample size $N_k = m_k + n_k$ to Welch's df for the two-sample comparison.

For all $K = 7$ studies a representative value of ξ is $\hat{\xi} = N/\sum_k \hat{\nu}_k$, where $N = \sum_k N_k$ is the total sample size. Note that $\hat{\xi}$ is a weighted harmonic mean (weights N_k/N) of the $\hat{\xi}_k$'s. For these data $N = 330$, $\sum_k \hat{\nu}_k = 307.9$ and $\hat{\xi} = 1.072$. The overall δ for the seven studies is defined by $\delta = \mathcal{K}_{\hat{\xi}}^{-1}(\kappa)$, and hence a 95 % confidence interval for δ is obtained by applying $\mathcal{K}_{\hat{\xi}}^{-1}$ to each endpoint of the 95 % confidence interval derived above for κ. The result is almost the same, [0.009,0.226], because the $\sinh^{-1}$ function behaves like the identity near the origin.

25.4.2 Inference for the random effects model

For the random (standardized) effects model the evidence for $\gamma > 0$ in $Q^* = 14.035$ is by definition (25.3) equal to $T'_{Q*} = 1.43$, which is very weak, so one should stay with the FEM analyzed above. But for the sake of illustration, we proceed with the analysis based on the REM.

The DerSimonian and Laird (1986) estimate (25.7) of the variance component γ^2 is $\hat{\gamma}^2 = 0.032$. For these data one can also compute $\bar{\kappa} = 0.162$, $s_\kappa = 0.2577$ and $\hat{\gamma}^2 = 0.030$ from (25.6).

The evidence for $\kappa > 0$ is by Definition 3 given by $T^*_{1:7} = 1.61$. Note that this is smaller than the evidence found earlier for $\kappa > 0$ using the fixed effects model. But there is not much difference because $\hat{\gamma}^2$ is small.

The 95 % Student t-interval (25.11) for κ is $[-0.077, 0.400]$. The 95 % interval for δ is readily found by applying the transformation $\mathcal{K}_{\hat{\xi}}^{-1}$ to each of the endpoints of the previous interval, which yields $[-0.076, 0.406]$. Note that these intervals are slightly larger than those obtained from κ, δ earlier, because they had to allow for a small variance component. If γ^2 were much larger, so would be these intervals.

25.5 Summary

In this chapter we have proposed methods for combining evidence in K studies for the fixed (equal or unequal) standardized effects model, as well as a random transformed standardized effects model. For all models, a representative standardized effect is defined and confidence intervals are provided. The methods are relatively simple because it is assumed that variance stabilization techniques have already transformed the test statistics onto the unit normal calibration scale. The reader is cautioned, however, that these techniques make strong assumptions, in particular that the evidence for each study is on the calibration scale to a good approximation, for all parameters of interest.

26

Correcting for publication bias

26.1 Publication bias

It is well known that the established practice of requiring experimental results to contradict a null hypothesis of no effect at level 0.05 introduces certain anomalies. The scientist who obtains a *p*-value of 0.049 may succeed in publishing the result, while the one who obtains 0.07 cannot publish. After reading this book it becomes clear that this is an absurd situation, because the second study contains almost as much evidence as the first one. The very fact of publication introduces a bias towards the alternative: a published *p*-value is *conditional* on its being less than a threshold. When combining *p*-values obtained through published studies, one must be aware of this selection bias and one must try to reduce its effect.

If selection bias affects a sample, one can sometimes make it visible in an appropriate plot. The missing parts show up as gaps, truncations, hollows, etc. In the literature on publication bias the funnel plot is often cited as such a tool.

Definition 26.1 *Suppose we plan to combine a group of similar studies. For each study two numerical summaries are at hand. First, an observed effect, which can be a log odds ratio, the deviation of a mean from the null value, or something else. Second, a measure of the precision, such as the standard error of the observed effect. The plot of the precision as a function of the effect is called a* ***funnel plot****.*

Example 26.1 *A simulated example may help in illustrating the selection introduced by publishing only studies that reach traditional significance as measured by a p-value of less than 0.05. In this example, we look at 300 studies, each resulting in an observed effect X that is normally distributed with a mean of μ and a standard*

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

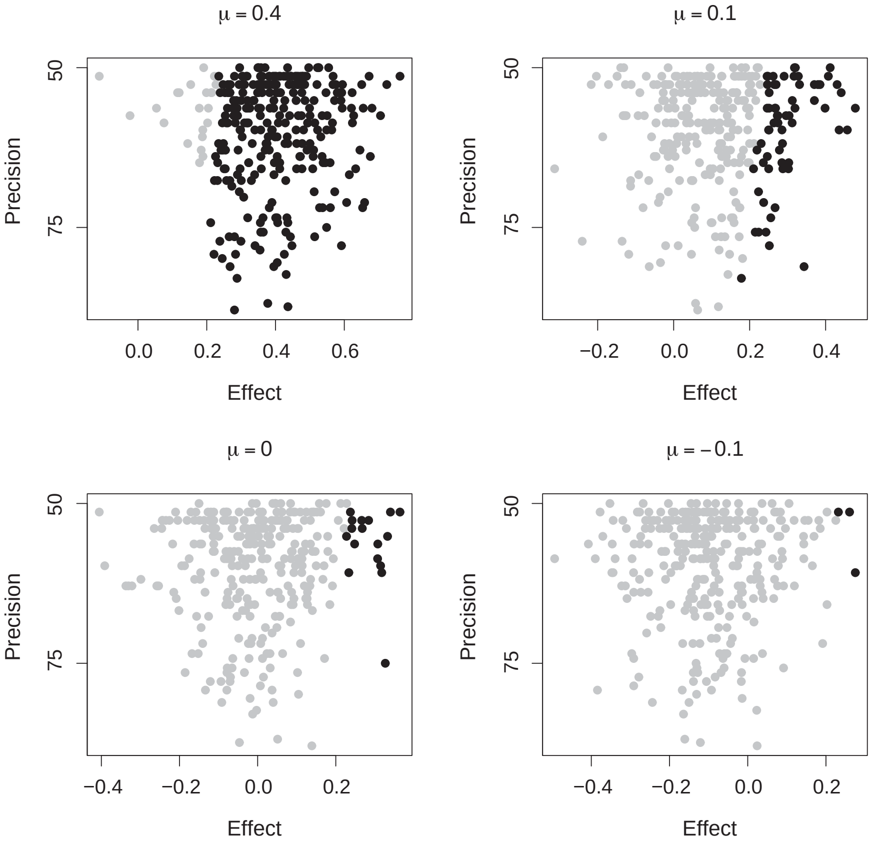


Figure 26.1 The four funnel plots show the precision (equal to the standard error) of each of 300 studies versus the observed effect X. Note that even though the study size is indicated in the y-axis labels, the ordinate used in plotting is $1/\sqrt{\text{study size}}$. A study rejects the null hypothesis, if the effect exceeds $1.645/\sqrt{\text{study size}}$. The dark points indicate the studies that reject the null. The four panels correspond to different actual effects. They go from $\mu = 0.4$ (upper left) to $\mu = 0.1$ (upper right) to $\mu = 0.0$ (lower left) and finally to $\mu = -0.1$ (lower right).

error of $1/\sqrt{n}$*, where* n *is the number of subjects in the study. Figure 26.1 shows what happens if we select the studies leading to a significant result, while ignoring the others. When the actual effect* μ *is large, or, more precisely, when the power of the study is close to one, the selection bias is negligible. This is the case in the upper left-hand panel of Figure 26.1. As the power decreases, the publication bias becomes more visible. In the upper right-hand panel, the funnel plot is asymmetrical and it is clear that more than one half of what should be there is missing.*

The most dangerous cases are shown in the lower row of plots, where the true effects are zero (no effect whatsoever) and -0.1*. In this latter case, the actual effect* *is in the opposite direction of the observed effect. In both cases, any published result is a false discovery. Imagine the funnel plot containing only the dark points. Would the reader, seeing these two plots, guess that something was amiss? In the lower left-hand plot one might convince oneself that the funnel plot is about half missing. In the right-hand plot, however, there are only three published studies and they confirm each other perfectly. And so, one would most likely conclude that a small positive effect truly existed.*

This example shows two things. For one, while the funnel plot is a valid idea in some circumstances, there are many ways in which things can go wrong. Using it as our tool for detecting bias is thus probably not a good idea. The second lesson is that unless we know more about how many unpublished studies have been performed, we cannot compute a reliable correction.

26.2 The truncated normal distribution

We argue in this book in favor of another presentation of the results from observational studies. Variance stabilizing the results of a study produces what we call *evidence* having variance about equal to one. The outcome of a study is then summarized in the evidence T, which is approximately normally distributed with mean $\sqrt{n}\kappa$ and variance 1. In our formulas we assume exact normality of the observed evidences.

The first model for publication bias we will consider is a conditional analysis of the published results. Being published implies that the p-value is below 0.05 or that the evidence obtained satisfies $t_i > 1.645$. Conditional on being published, this means that the observed result t_i no longer has a normal distribution, but rather a truncated normal distribution, because t_i is guaranteed to exceed a certain bound.

Definition 26.2 *A random variable X is said to have a truncated normal distribution with truncation point β ($X \sim \mathcal{TN}(\mu, \sigma^2, \beta)$) if it has density*

$$f(x|\mu,\sigma,\beta) = \frac{\varphi((x-\mu)/\sigma)/\sigma}{1-\Phi((\beta-\mu)/\sigma)} \quad \textit{for } x \geq \beta\,. \tag{26.1}$$

For $x < \beta$, the density is equal to zero. The parameters of this distribution are μ, σ and β.

The mean of a truncated normal distribution (26.1) is equal to

$$\mathrm{E}[X] = \mu + \frac{\sigma\,\varphi((\beta-\mu)/\sigma)}{1-\Phi((\beta-\mu)/\sigma)},$$

which contains an expression for the numerical size of the bias one incurs when using X for estimating μ.

The published evidence T_i of a study has a very simple truncated normal distribution

$$T_i \sim \mathcal{TN}(\sqrt{n_i}\kappa, 1, 1.645),$$

and assuming that $t_1, \ldots, t_m$ are a sample with these distributions, the likelihood for κ is

$$L_{\text{truncated}}(\kappa) = \prod_{i=1}^{m} f(t_i|\kappa\sqrt{n_i}, 1, 1.645)\,. \tag{26.2}$$

An algorithm for maximizing the likelihood can be based on our expression for the mean of a truncated normal. At the start, we simply ignore the bias and treat t_i as if it were normally distributed. The estimate of κ is then

$$\widehat{\kappa} = \sum_{i=1}^{m} \sqrt{n_i}\, t_i \Big/ \sum_{i=1}^{m} n_i.$$

Based on this, we can estimate by how much we overestimate the true effect κ and make a correction. The resulting algorithm is as follows:

1. Put $k = 0$ and
$$\widehat{\kappa}_k = \sum_{i=1}^{m} \sqrt{n_i}\, t_i \Big/ \sum_{i=1}^{m} n_i.$$
2. Compute the corrections
$$b_i = \frac{\varphi(1.645 - \widehat{\kappa}_k\sqrt{n_i})}{1 - \Phi(1.645 - \widehat{\kappa}_k\sqrt{n_i})}$$
for $i = 1, \ldots, m$.
3. Update the estimate by putting $k = k + 1$ and
$$\widehat{\kappa}_k = \sum_{i=1}^{m} \sqrt{n_i}\,(t_i - b_i) \Big/ \sum_{i=1}^{m} n_i.$$
4. Stop the calculations and put $\widehat{\kappa}_{\text{truncated}}$ equal to the final value, as soon as the estimate does not change any more, otherwise return to step 2.

Once the estimate of the underlying effect κ is obtained, we have corrected for the publication bias. The combined evidence is estimated by

$$T_{\text{combined by truncation}} = \sum_{i=1}^{m} \sqrt{n_i}\, \widehat{\kappa}_{\text{truncated}} / \sqrt{m},$$

Table 26.1 Data relating to Example 26.2.

k	$\widehat{\kappa}_k$
0	0.5
1	0.35
2	0.26
5	0.11
10	0.00
20	–0.09
30	–0.11
40	–0.12

and the combined p-value, corrected for publication bias, is

$$p_{\text{combined by truncation}} = 1 - \Phi(T_{\text{combined by truncation}}).$$

Example 26.2 *Table 26.1 shows what happens when applying this procedure to the case of the one smallish study of* $n = 16$ *subjects that gave significant evidence of* $t = 2$*. The estimated bias is considerable. While the naive use of the lone published study predicts that* $\widehat{\kappa} = 0.5$*, the conditional analysis based on the truncation model predicts a value of* $\widehat{\kappa}_{\text{truncated}} = -0.12$*, that is, an actual effect in the opposite direction. As a consequence,* $p_{\text{combined by truncation}} = 0.68$ *is larger than 0.5 and the evidence* $T_{\text{combined by truncation}} = -0.48$ *is negative.*

26.3 Bias correction based on censoring

The estimation by the truncated normal is usually not quite the right thing to do, because it provides the same correction, whether there were any unpublished studies or not. Intuitively one would think, however, that the number of unpublished studies ought to play a part. If a single study is done and it has a p-value that is smaller than 0.05, why should one correct for bias? If, on the other hand, only one in 300 studies results in such evidence, why should one believe its result? The interpretation of published evidence is completely different depending on whether the published study is the only one ever done, or whether it is the only one with an observed effect that reaches the standard of traditional significance among many studies. In the first case, we would say that no bias is present, whereas in the second case a vigorous bias correction is needed.

Why does the conditional model we described above, and which seems plausible, sometimes fail? Well, it assumes that the researchers performing the studies will continue repeating them until one reaches a result with an associated p-value smaller than 0.05. In this way, one is guaranteed to obtain a published study and in effect the truncated normal is the correct model. In this model, each published study has its natural proportion of accompanying unpublished studies. In reality, though, studies are planned and performed without the intention of repeating them until one has a sufficiently large observed effect. The evidence obtained by each study is thus not modeled by the truncated normal, but rather by the normal itself. The bias is still there, but the reason is not truncation, it is rather censoring. The evidence of any study that happens not to reach the required p-value of 0.05 is suppressed.

If we had more detailed information about the unpublished studies – how many there were and what sample sizes were used – we could take it into account by replacing the previous likelihood (26.2) by

$$L_{\text{censored}}(\kappa) = \prod_{i=1}^{m} \varphi(t_i - \sqrt{n_i}\kappa) \prod_{j=1}^{l} \Phi(1.645 - \kappa\sqrt{n_j^*}). \tag{26.3}$$

Here the number of unpublished studies is equal to l and the sample sizes are $n_1^*, \ldots, n_l^*$. The published studies are characterized by the evidences $t_1, \ldots, t_m$ and sizes $n_1, \ldots, n_m$. The maximum likelihood estimate of κ satisfies

$$\sum_{i=1}^{m} \sqrt{n_i}\, \left(t_i - \sqrt{n_i}\,\widehat{\kappa}_{\text{censored}}\right) - \sum_{j=}^{l} \sqrt{n_j^*} \frac{\varphi(1.645 - \sqrt{n_j^*}\,\widehat{\kappa}_{\text{censored}})}{\Phi(1.645 - \sqrt{n_j^*}\,\widehat{\kappa}_{\text{censored}})} = 0. \tag{26.4}$$

Of course, we do not know how many unpublished studies have been performed and, in order to use the new likelihood, we need to decide how big to choose l, the number of unpublished studies and what to take for $n_1^*, \ldots, n_l^*$. The second choice is the easier one. We propose to use the average size of the published studies, that is, to put $n_j^* = \sum_{i=1}^{m} n_i/m$ for all j. As for the number of unpublished studies, we propose to compute the bias corrected evidence and p-values for a variety of choices and leave it to the user to make the final decision.

An algorithm that works reasonably well for solving the likelihood equation is the Newton–Raphson iteration. It leads to the following little program.

1. Put $k = 0$ and

$$\widehat{\kappa}_k = \sum_{i=1}^{m} \sqrt{n_i}\, t_i \Big/ \sum_{i=1}^{m} n_i.$$

For each of the l presumed latent or unpublished studies, set $n_j^* = \bar{n}$, the average study size of the observed studies.

2. Compute for $j = 1, \ldots, l$, that is, for the latent studies, the quantity $u_j = 1.645 - \sqrt{n_j^*}\hat{\kappa}_k$.

(a) Compute the log-likelihood derivative

$$f = \sum_{i=1}^{m} \sqrt{n_i}(t_i - \sqrt{n_i}\widehat{\kappa}) - \sum_{j=1}^{l} \sqrt{n_j^*}\,\varphi(u_j)/\Phi(u_j).$$

(b) Compute the second derivative of the log-likelihood

$$f' = -\sum_{i=1}^{m} n_i - \sum_{j=1}^{l} N_j^* \,(\varphi(u_j)/\Phi(u_j)\,(u_j + \varphi(u_j)/\Phi(u_j)).$$

3. Update the estimate by putting $k = k + 1$ and

$$\widehat{\kappa}_\kappa = \widehat{\kappa}_{k-1} - f/f'.$$

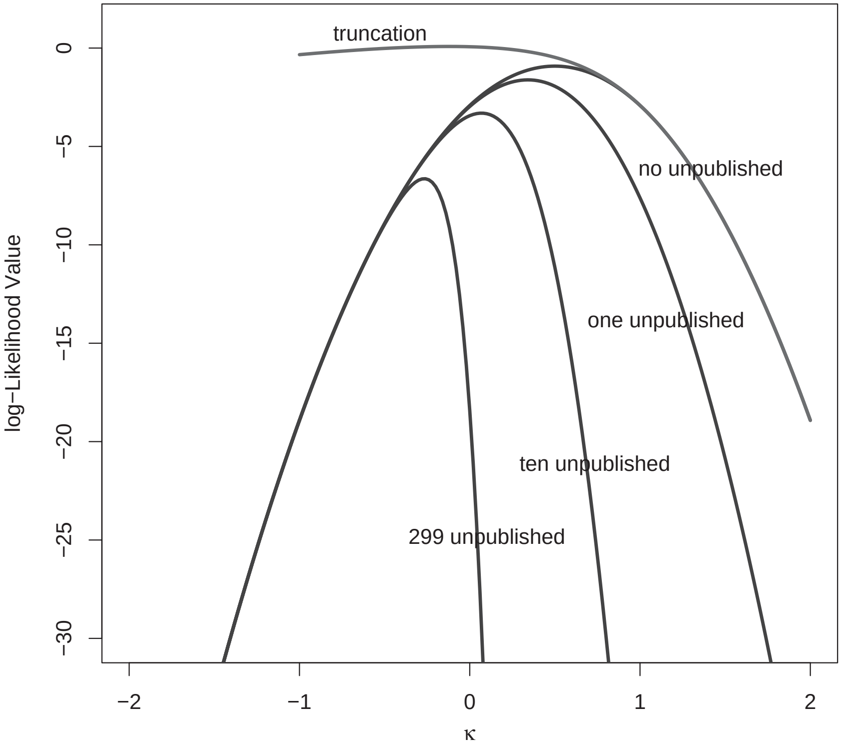


Figure 26.2 The figure shows the log-likelihood function for κ for the data of Example 26.3. When we assume that the single available study is the only one ever performed (none unpublished), then the maximum likelihood estimate is $\widehat{\kappa} = 2/4 = 0.5$. We have seen previously that for the truncation model, $\widehat{\kappa}_{\text{truncated}} = -0.12$. For the censored case the values at which the various curves attain their maxima yield the corresponding estimates. These values are equal to $\widehat{\kappa}_{\text{censored}} = 0.34\,, 0.07\,, -0.26$ for $l = 1\,, 10\,, 299$, respectively.

4. Stop the calculations and put $\widehat{\mu}_{\text{censored}}$ equal to the final value, as soon as the estimate does not change any more, otherwise return to step 2.

We could simplify the formulas somewhat by using the fact that in our proposed procedure all the n_j^* have the same value, but we chose not to do so in order to give the algorithm in full generality.

Example 26.3 *To illustrate the behavior of the censored and truncated log-likelihood functions, consider again the example where a single study is available. It has weak evidence of* $t = 2$ *and the number of subjects used was* $n = 16$. *Figure 26.2 shows various likelihood functions for the underlying effect* κ. *If we assume that one other study of similar size has been left unpublished, the combined evidence is* $T_{\text{combined by censoring}} = 1.37$ *and the corrected* p*-value becomes* $p_{\text{combined by censoring}} = 0.085$. *For the censored case, we assume that a number of unpublished studies, not attaining a p-value of 0.05, had been performed. The size of these unpublished studies is taken to be equal to* $n = 16$, *the size of the published study.*

26.4 Summary

Biasing findings by selectively publishing only those studies that reach a certain standard, while suppressing those that do not, has been called *publication bias*. In this chapter we have shown two simple ways in which one can combine the results from several studies and correct for this bias. The first method is based on a truncation model. It usually results in quite a vigourous correction, but has the advantage of not requiring any knowledge beyond the results of the published studies.

The second method is based on censoring. To implement it, we need to know the number of unpublished studies as well as the number of subjects used in each of the unpublished studies. In other words, some information about the unobserved latent data must be available. For this method, we make a practical proposal that only requires the user to guess the number of unpublished studies.

In the litterature on publication bias, the funnel plot is often advocated as a tool for detecting the bias and even correcting for it. In our opinion, however, this is not a safe method and we do not recommend its use.

For further reading on these topics, we invite the reader to consult Chapter 15 at `http://www.cochrane-net.org/openlearning` as well as the article by Givens *et al.* (1997).

27

Large-sample properties of variance stabilizing transformations

27.1 Existence of the variance stabilizing transformation

The following description of (asymptotic) univariate variance stabilizing transformations (*vst*'s) is taken from Holland (1973), which gives a nicely written account of the subject.

Let X_n be a real-valued random variable with distribution depending upon a real parameter $\theta \in D$, an open interval in $\mathbb{R}$. X_n may for example be an estimator based on a sample of size n. Suppose that for every $\theta \in D$ the quantity $\sqrt{n}(X_n - \theta) \rightarrow N(0, \sigma^2(\theta))$ in distribution. The asymptotic variance $\sigma^2(\theta) > 0$ is assumed to be continuous in D.

An asymptotic *vst* is a one-to-one, continuously differentiable mapping $f : D \rightarrow \mathbb{R}^1$ such that $\sqrt{n}(f(X_n) - f(\theta)) \rightarrow N(0, 1)$ in distribution. Since $X_n \rightarrow \theta$ in probability, $X_n \in D$ with probability as close to 1 as needed for n large enough. Therefore f is defined for the possible values of X_n with a probability approaching 1 as $n \rightarrow \infty$. This situation is, of course, not satisfactory in practice. As a remedy and in order to apply a *vst*, one may have to extend the definition of $f(\cdot)$.

Assume that f exists and has a differential at each $\theta \in D$, i.e. if $|x_n - \theta| = O(n^{-1/2})$ then $f(x_n) = f(\theta) + (x_n - \theta)f'(\theta) + o(n^{-1/2})$. For the random variable

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephan Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

X_n the same is true in probability, resulting in

$$\sqrt{n}(f(X_n) - f(\theta)) \to N(0, \sigma^2(\theta)(f'(\theta))^2)$$

in distribution. Since only continuously differentiable solutions of the differential equation $\sigma^2(\theta)(f'(\theta))^2 = 1$ are acceptable the sign should be 1 or -1 for all $\theta \in D$. In summary, the one-dimensional asymptotic *vst* problem always has a one-to-one continuously differentiable solution given by

$$f(\theta) = f(\theta_0) \pm \int_{\theta_0}^{\theta} (\sigma(t))^{-1} \mathrm{d}t. \tag{27.1}$$

The solution is unique up to an additive constant and the sign of its derivative. The only requirement is that $\sigma(\theta)$ is a continuous nonzero function of θ in D, it does not have to be one-to-one and may be constant. In the following, we will choose the positive sign in the defining equation for the *vst*. The additive constant allows us to fix the value of the *vst* at one point, for example, $f(\theta_0) = 0$.

So far we were looking at the asymptotic *vst* valid in the $n^{-1/2}$-vicinity of θ_0. But often the *vst* is defined on a much larger region, or even globally, as will be seen in the examples in the next section. Interestingly, in all these examples, and in the majority of variance functions for traditional exponential families, the variance $\sigma^2(\theta)$ is a first- or second-degree polynomial in θ. In such cases a global *vst* exists and is a rather simple transformation.

We also were working in a most simple case of $\sigma^2 = \sigma^2(\theta)$, but a more general case is $\sigma^2 = \sigma^2(\xi; \theta)$, where ξ is a nuisance parameter. An example is the *vst* for the Student t, where the variance is the nuisance parameter. The presence of nuisance parameters modifies the above asymptotic theory as follows. Equation (27.1) changes to

$$f(\theta|\xi) = f(\theta_0|\xi) + \int_{\theta_0}^{\theta} (\sigma(\xi; t))^{-1} \mathrm{d}t, \tag{27.2}$$

which means that the *vst* depends on ξ. Suppose $\hat{\xi}$ is asymptotically independent of $X_n = \hat{\theta}$, and $\hat{\xi} \to \xi$ in probability, then we may solve (27.2) with $\hat{\xi}$ substituted for ξ.

The asymptotic *vst* is based on the asymptotic variance $\sigma^2(\theta)$ and Equation (27.1). In a finite sample setting, approximate variance stabilization can be achieved by applying (27.1) to the actual variance $\sigma_n^2(\theta)$. We denote the finite sample *vst* by $f_n(\cdot)$. When $n \to \infty$, the *vst* has the effect of rendering the asymptotic variance equal to 1 for all θ. For finite n, this holds only approximately, but in practice often goes a long way towards this goal.

27.2 Tests and effect sizes

Let us now compare tests and effect sizes before and after the *vst*. In the previous section we considered an estimator $X_n = \hat{\theta}$. Its mean $\mathrm{E}(X_n) = \theta$ was the main parameter of interest. In a slightly more general setting we shall consider test statistics X_n of a null hypothesis involving a real-valued parameter ζ. Denote the expectation of the test statistic $\theta(\zeta) = \mathrm{E}_\zeta(X_n)$. Assume without loss of generality that under the null hypothesis $\zeta = 0$ and let $\theta_0 = \theta(0)$. The effect size associated with the test based on $Y_n = \sqrt{n}\,(X_n - \theta_0)$ is

$$\delta = (\theta - \theta_0)/\sigma(\theta) = (\theta - \theta_0) f'(\theta), \tag{27.3}$$

where $f(\cdot)$ is the asymptotic *vst* for X_n.

The Pitman efficacy of a test describes the behavior of the asymptotic power. The Pitman efficacy of the test Y_n is

$$e_Y = (d\theta(0)/d\zeta)\,\sigma(\theta_0)^{-1} = (d\theta(0)/d\zeta)\, f'(\theta_0).$$

This result holds if $\theta(\zeta)$ is differentiable in ζ at 0 with a positive derivative and σ is continuous at θ_0 and nonzero. The ARE of such tests is the ratio of squared efficacies, see Theorem 14.19 from van der Vaart (1998).

After application of the *vst* we obtain what we call evidence statistic throughout this book. This is another test, which has the form $T_n = \sqrt{n} f(X_n)$. For T_n we have weak convergence to a unit normal distribution $\sqrt{n}(f(X_n) - f(\theta)) \to N(0, 1)$. This new test statistic thus has an effect size of

$$f(\theta) - f(\theta_0). \tag{27.4}$$

It is easy to see that the Pitman efficacy of a test is not affected by the application of a *vst*. The two tests based on Y_n and T_n are asymptotically equivalent.

Lemma 27.1 *The Pitman efficacy of a test remains constant under the application of the variance stabilizing transformation. In this sense, the tests Y_n and T_n are equivalent.*

The proof is straightforward. The efficacy of T_n is computed with the help of (27.4) and equals $e_T = \mathrm{d} f(\theta)/\mathrm{d}\zeta$, where the derivative is evaluated at $\zeta = 0$. The chain rule then leads to $e_T = (\mathrm{d}\theta(0)/\mathrm{d}\zeta)\, f'(\theta_0) = e_Y$.

Comparing the original effect size δ and the effect size after variance stabilization $f(\theta) - f(\theta_0)$ on an interval (θ_0, θ) we obtain the following result.

Lemma 27.2 *Suppose the vst $f(\cdot)$ is twice continuously differentiable. It follows that the effect size of the transformed test T_n is larger than the effect size of the original test Y_n if and only if (iff) the vst is concave on (θ_0, θ). This holds iff $\mathrm{d}\sigma/\mathrm{d}\theta > 0$, which means that σ is an increasing function of the parameter θ.*

Proof: We expand $f(\theta_0)$ around the θ and obtain $f(\theta_0) = f(\theta) + f'(\theta)(\theta_0 - \theta) + \{f''(c)/2\}\,(\theta - \theta_0)^2$, for some c lying between θ_0 and θ. When applying this to the effect size for the test based on T_n we have

$$\begin{aligned} f(\theta) - f(\theta_0) &= (\theta - \theta_0) f'(\theta) - \frac{(\theta - \theta_0)^2}{2} f''(c) \\ &= \delta - \frac{(\theta - \theta_0)^2}{2} f''(c). \end{aligned}$$

The transformed effect size is thus larger than the original effect size iff the *vst* is concave on (θ_0, θ), i.e. iff $f''(\theta) < 0$.

Recall that $f'(\theta) = \sigma(\theta)^{-1}$. From this it follows that $f''(\theta) = -(\sigma(\theta))^{-2}\mathrm{d}\sigma/\mathrm{d}\theta$. This shows that the *vst* $f(\cdot)$ is concave on (θ_0, θ) iff $\mathrm{d}\sigma/\mathrm{d}\theta > 0$.

Example 1. Poisson counts

We observe a sample of counts, each having a Poisson distribution with expectation μ. The estimate for μ is the sample mean X_n, which satisfies $\sqrt{n}(X_n - \mu) \to N(0, \mu)$. To test $\mu = \mu_0$ versus $\mu > \mu_0$ we use $Y_n = \sqrt{n}(X_n - \mu_0)$. The asymptotic *vst* for X_n is up to an additive constant and a sign change equal to twice the square root, so that $T_n = \sqrt{n} f(X_n) = 2\sqrt{Y_n}$.

The effect size before applying the *vst* is $\delta = (\mu - \mu_0)/\sqrt{\mu}$. The transformed effect is $2(\sqrt{\mu} - \sqrt{\mu_0})$. The derivative of $\sigma(\mu)$ is positive for $\mu > 0$, therefore the *vst* increases the effect size.

Example 2. The *t*-test

The parameter of interest when using the t-statistic is the mean μ, but we are in the presence of a nuisance parameter, the variance σ^2. The test is constructed with the help of X_n and s_n, the sample mean and standard deviation, which are asymptotically independent. We reject $\mu = \mu_0$ in favor of $\mu > \mu_0$ for large values of $Y_n = \sqrt{n}(X_n - \mu_0)/s_n$. A relevant standardized effect is Cohen's $d = (\mu - \mu_0)/\sigma$, and $Y_n = \sqrt{n}\hat{d}_n$. The statistic Y_n is approximately normal with mean $\sqrt{n}d$ and variance $\sigma^2(d) = 1 + d^2/2$ and the corresponding finite sample *vst* is discussed in Chapter 20. One finds

$$f_n(\hat{d}_n) = \sqrt{2}\,\ln\left(\hat{d}_n/\sqrt{2} + \sqrt{1 + (\hat{d}_n/\sqrt{2})^2}\right).$$

The effect size before applying the *vst* is $\delta = d/\sqrt{1 + d^2/2}$. For the statistic $T_n = \sqrt{n} f_n(\hat{d}_n)$, the effect size is $f(d) = \sqrt{2}\ln(d/\sqrt{2} + \sqrt{1 + d^2/2})$. The two effect sizes are very close for small values of d. Say for $d = 0.05$, $\delta = 0.049967$ and $f(d) = 0.04999$. These functions grow very far apart for large values of d, with $\delta \to \sqrt{2}$ in the limit, whereas $f(\delta)$ is unlimited.

Example 3. Binomial proportions

Here $\sigma(p) = \sqrt{p(1-p)}$, and $\delta = (p - p_0)/\sigma(p)$. The sign $\mathrm{sgn}(\mathrm{d}\sigma/\mathrm{d}\theta) = \mathrm{sgn}((1/2) - p)$. The sign is constant on (p_0, p) if p_0 and p are at the same side of 1/2. When $p_0 < p < 0.5$ the *vst* should result in increased effect size. When $0.5 < p_0 < p$ the transformed effect size should be smaller. The transformed effect size is

$$f(\delta) = \arcsin(1 - 2p_0) - \arcsin\left(\frac{1-2p_0-\delta\sqrt{1+\delta^2-(1-2p_0)^2}}{1+\delta^2}\right). \qquad (27.5)$$

When $p_0 = 1/2$,

$$f(\delta) = \arcsin\left(\frac{\delta}{\sqrt{1+\delta^2}}\right) = \arctan(\delta).$$

This shows that $|f(\delta)| \leq |\delta|$ for $p_0 = \frac{1}{2}$.

Example 4. The sign test

Given n observations from a continuous distribution $F(\mu, \sigma)$ with unknown median μ and scale parameter σ, we wish to test $\mu = 0$ in favor of $\mu > 0$. The sign statistic $S_n = \sum_i I\{X_i > 0\} \sim$ Binomial (n, p_μ), with $p_\mu = 1 - F(-\mu/\sigma)$. This is exactly the previous example with $p_0 = 1/2$ and $p_\mu > 1/2$. The effect size δ is positive, and it is decreased by the *vst*.

27.3 Power and efficiency

The power of an asymptotically normal α-level test based on Y_n is approximately equal to $1 - \Phi(z_{1-\alpha} - \sqrt{n}\delta)$, where δ is given by (27.3). The sample size of a test with power $1 - \beta$ can be calculated from

$$\sqrt{n_Y} = \delta^{-1}(z_{1-\alpha} + z_{1-\beta}). \tag{27.6}$$

Similarly, after the *vst*, the test based on T_n has power $1 - \Phi(z_{1-\alpha} - \sqrt{n}(f(\theta) - f(\theta_0))$, and the sample size is calculated from

$$\sqrt{n_T} = (f(\theta) - f(\theta_0))^{-1}(z_{1-\alpha} + z_{1-\beta}). \tag{27.7}$$

Note that (27.6) implies that $\delta = (z_{1-\alpha} + z_{1-\beta})/\sqrt{n_Y}$, which is small for large sample sizes. Recall from Lemma 27.1 that for small values of δ we found $f(\theta) = f(\theta_0) + (\theta - \theta_0)\, f'(0) + o(\delta)$, which means that for large sample sizes (27.6) and (27.7) give approximately identical sample sizes, since the above implies that $f(\theta) - f(\theta_0) \approx \delta$. In many practical cases, though, this asymptotic equivalence is not sufficiently accurate and the two sample sizes n_Y and n_T may be quite different. The ratio of (nominal) sample sizes is

$$n_Y/n_T = ((f(\theta) - f(\theta_0))/\delta)^2. \tag{27.8}$$

The following result is a corollary of Lemma 27.2.

Corollary 27.3 *Asymptotic sample size calculation based on Y_n results in a larger/smaller sample size than the one based on T_n (i.e. $n_Y > n_T$) iff the vst is concave ($\mathrm{d}\sigma/\mathrm{d}\theta > 0$)/convex ($\mathrm{d}\sigma/\mathrm{d}\theta < 0$).*

Example 5. The *t*-test (continued)

The effect size of the t-statistic is $\delta = d/\sqrt{1 + d^2/2}$ and becomes $\sqrt{2}\ln(d/\sqrt{2} + \sqrt{1 + d^2/2})$ after the stabilizing transformation, where $d = (\mu - \mu_0)/\sigma$ specifies the alternative. The ratio of the nominal sample sizes is

$$n_Y/n_T = (2+d^2)d^{-2}[\ln(d/\sqrt{2}+\sqrt{1+d^2/2})]^2.$$

Fixing the sample size n and the false positive and false negative error rates α and β, one can solve Equation (27.6) for d. Substituting the result in the above equation, we obtain an expression for the ratio of nominal sample sizes n_Y/n_T for a given sample size, level and power, denoted by $r(Y, T|n, \alpha, \beta)$. We leave it to the reader to check that this leads to

$$r(Y, T|n, \alpha, \beta) = \frac{2n}{(z_{1-\alpha}+z_{1-\beta})^2}\left[\ln\frac{(z_{1-\alpha}+z_{1-\beta})+\sqrt{2n}}{\sqrt{2n-(z_{1-\alpha}+z_{1-\beta})^2}}\right]^2. \qquad (27.9)$$

This ratio of sample sizes is a decreasing function of all three parameters, the sample size and the false positive and false negative error rates α and β. The limit of $r(Y, T|n, \alpha, \beta)$ when the sample size $n \to \infty$ is 1, but for moderate values of n (between 10 and 100) the ratio is considerably greater than 1. Some examples are given in Table 27.1.

Example 6. The sign test (continued)

Consider testing $H_0 : p_\mu = 1 - F(-\mu/\sigma) = 1/2$ versus $H_A : p_\mu > 1/2$. The effect size of the sign statistic is $\delta = (p_\mu - 1/2)/\sqrt{p_\mu(1-p_\mu)}$. After transformation to evidence via the *vst*, the effect size is $f(\delta) = \arctan(\delta)$. The ratio of the sample sizes is $n_Y/n_T = (\delta/\arctan(\delta))^2$. Substituting $\delta = (z_{1-\alpha}+z_{1-\beta})/\sqrt{n}$ in the above equation, we obtain an expression for the ratio $r(Y, T|n, \alpha, \beta)$ for a given triplet (n, α, β). This is an increasing function of the error rates α and β, and of the sample size n. The limit when the sample size $n \to \infty$ is 1, but for moderate values of n (between 10 and 100) the ratio $r(Y, T|n, \alpha, \beta)$ is considerably smaller than 1, i.e. the asymptotic

Table 27.1 Values of the ratio of nominal sample sizes $r(Y, T|n, \alpha, \beta)$ calculated from Equation (27.9) for *t*-test (columns 2–4) and the sign test (columns 5–7; see the text for explanation) for $\alpha = 0.05$, $\beta = 0.05$, 0.10 and 0.20 and for various sample sizes n.

	t-test			Sign test		
n	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.2$	$\beta = 0.1$	$\beta = 0.05$
10	1.27	1.43	1.64	0.72	0.65	0.60
15	1.16	1.24	1.34	0.79	0.73	0.69
20	1.12	1.17	1.23	0.83	0.78	0.74
25	1.09	1.13	1.17	0.86	0.82	0.78
30	1.07	1.11	1.14	0.88	0.84	0.81
35	1.06	1.09	1.12	0.90	0.86	0.83
40	1.05	1.08	1.10	0.91	0.88	0.85
45	1.05	1.07	1.09	0.92	0.89	0.86
50	1.04	1.06	1.08	0.92	0.90	0.88
100	1.02	1.03	1.04	0.96	0.95	0.93

Table 27.2 Comparison of sample sizes for the t-test at $\alpha = 0.05$, $\beta = 0.05, 0.10$ and 0.20. The sample sizes were selected for the t-test (n_t) and calculated using Equation (27.7) for the evidence-based test (n_T). To do so, the value of $\delta_t = (\mu - \mu_0)/\sigma$ was calculated with the help of the NCSS-PASS (2005) software in order to match the chosen n_t, α and β.

n_t	β	δ_t	$f(\delta_t)$	n_T	n_t/n_T
10	0.05	1.131	1.036	10.09	0.99
15	0.05	0.894	0.843	15.22	0.99
20	0.05	0.764	0.731	20.25	0.99
25	0.05	0.677	0.653	25.34	0.99
30	0.05	0.615	0.597	30.35	0.99
35	0.05	0.568	0.554	35.29	0.99
40	0.05	0.529	0.517	40.43	0.99
45	0.05	0.498	0.488	45.40	0.99
50	0.05	0.472	0.464	50.34	0.99
100	0.05	0.331	0.328	100.56	0.99
10	0.1	1.005	0.935	9.79	1.02
15	0.1	0.795	0.758	14.90	1.01
20	0.1	0.679	0.655	19.94	1.00
25	0.1	0.603	0.586	24.93	1.00
30	0.1	0.547	0.534	30.01	1.00
35	0.1	0.505	0.495	34.97	1.00
40	0.1	0.471	0.463	40.00	1.00
45	0.1	0.443	0.436	45.04	1.00
50	0.1	0.420	0.414	49.95	1.00
100	0.1	0.295	0.293	99.82	1.00
10	0.2	0.853	0.808	9.46	1.06
15	0.2	0.675	0.652	14.56	1.03
20	0.2	0.577	0.562	19.57	1.02
25	0.2	0.512	0.501	24.59	1.02
30	0.2	0.465	0.457	29.60	1.01
35	0.2	0.429	0.423	34.61	1.01
40	0.2	0.400	0.395	39.66	1.01
45	0.2	0.376	0.372	44.75	1.01
50	0.2	0.357	0.353	49.53	1.01
100	0.2	0.250	0.249	99.94	1.00

sample size calculation based on the standard normal approximation to the sign test results in a considerably smaller sample size n_Y in comparison to the evidence-based sample size calculation n_T. Some examples are given in the last three columns of Table 27.1.

The results of these two examples are rather striking. The evidence-based sample size calculations for the t-test for sample sizes up to 100 give considerably smaller values of n, whereas for the sign test they result in considerably larger values of

Table 27.3 Comparison of sample sizes for the sign test at nominal level $\alpha = 0.05$ and for three values of the type II error β. The sample size n_S was chosen, whereas n_Y and n_T were calculated with the help of p_μ, the alternative corresponding to the triplet (n_S, α, β). These were computed by the program NCSS-PASS (2005).

n_S	p_μ	α	β	δ	n_T	n_Y	$n_T - n_S$	$n_S - n_Y$
10	0.963	0.011	0.05	2.460	11.08	2.57	1.08	7.43
15	0.903	0.018	0.05	1.365	15.98	7.56	0.98	7.44
20	0.860	0.021	0.05	1.040	20.95	12.55	0.95	7.45
25	0.830	0.022	0.05	0.877	25.92	17.47	0.92	7.53
30	0.779	0.049	0.05	0.672	31.02	24.05	1.02	5.95
35	0.764	0.045	0.05	0.623	35.99	28.78	0.99	6.22
40	0.753	0.040	0.05	0.586	40.97	33.53	0.97	6.47
45	0.743	0.036	0.05	0.556	45.93	38.25	0.93	6.75
50	0.735	0.032	0.05	0.532	50.95	43.03	0.95	6.97
100	0.664	0.044	0.05	0.346	100.98	93.59	0.98	6.41
10	0.946	0.011	0.1	1.963	10.61	3.33	0.61	6.67
15	0.878	0.018	0.1	1.156	15.60	8.58	0.60	6.42
20	0.834	0.021	0.1	0.898	20.60	13.67	0.60	6.33
25	0.804	0.022	0.1	0.765	25.58	18.64	0.58	6.36
30	0.752	0.049	0.1	0.585	30.71	25.15	0.71	4.85
35	0.739	0.045	0.1	0.544	35.71	29.94	0.71	5.06
40	0.729	0.040	0.1	0.514	40.67	34.70	0.67	5.30
45	0.720	0.036	0.1	0.490	45.64	39.46	0.64	5.54
50	0.713	0.032	0.1	0.470	50.61	44.22	0.61	5.78
100	0.647	0.044	0.1	0.306	100.72	94.83	0.72	5.17
10	0.917	0.011	0.2	1.508	10.16	4.34	0.16	5.66
15	0.843	0.018	0.2	0.942	15.23	9.80	0.23	5.20
20	0.799	0.021	0.2	0.745	20.25	14.96	0.25	5.04
25	0.770	0.022	0.2	0.641	25.25	19.97	0.25	5.03
30	0.718	0.049	0.2	0.486	30.41	26.35	0.41	3.65
35	0.707	0.045	0.2	0.455	35.39	31.18	0.39	3.82
40	0.698	0.040	0.2	0.432	40.37	35.98	0.37	4.02
45	0.691	0.036	0.2	0.413	45.34	40.78	0.34	4.22
50	0.685	0.032	0.2	0.398	50.34	45.60	0.34	4.40
100	0.626	0.044	0.2	0.260	100.41	96.12	0.41	3.88

n than the simple asymptotic approximation of the traditional test statistic would lead one to believe. How do the *vst*-based sample sizes compare with exact sample sizes obtained from the noncentral t or from the binomial distribution? The NCSS-PASS[1] (2005) software was used to obtain the values of $\delta_t = (\mu - \mu_0)/\sigma$

[1] Power analysis and sample size software produced by NCSS (http://www.ncss.com)

(effect size for t-test) for given values of n, α and β (see Table 27.2). Then the *vst* was applied to calculate the effect size of the evidence-based t-test, $f(\delta_t)$. Finally, the sample size n_T was calculated with (27.7). Surprisingly, the sample sizes agree down to $n = 10$, which shows that sample size computations based on the evidence are very accurate. Note that minor differences are explained by using exact and not rounded-up sample sizes for n_T.

For the sign test at nominal level $\alpha = 0.05$ and fixed β, the values of the actual levels α and P_μ (probability under alternative) were calculated by the NCSS-PASS (2005) for a given sample size n_S. Then effect size δ (effect size for sign test) and transformed effect size $f(\delta)$ were calculated as in Example 4, and were used to calculate approximate sample sizes n_Y and n_T using not the nominal, but the true α level. Table 27.3 contains the numerical values. The differences between the calculated sample sizes and the actual sample sizes derived from the program are given in last two columns. The evidence-based sample size is within 1 of the true sample size, whereas the classic asymptotic sample size calculation substantially underestimates the sample size needed.

27.4 Summary

In this chapter we have seen that under an assumption of asymptotic normality and some standard regularity conditions, the *vst* always exists. Evidence obtained via a *vst* is also asymptotically normal. Its ARE to the original test is 1. We have also demonstrated that the *vst* may both increase and decrease (positive) effect size, depending on the behavior of variance as the function of the distance from the null. This difference of the effect sizes may be very large, even unlimited. When the variance increases, the *vst* increases the effect size. When the opposite is true, the variance is the highest at the null (see Examples 3 and 4 above). In this case the effect size decreases when stabilizing the distribution. Finally, sample size calculations based on variance stabilizing transformations perform considerably better than standard asymptotic sample size calculations for sample sizes up to 100, as was shown for the t-test and the sign test.

Acknowledgements

The authors would like to thank the Editors at Wiley Interscience, in particular Kathryn Sharples and Beth Dufour, for their advice and assistance in the production of this book. The software packages R and MINITAB were employed in simulation studies. The plots in Figures 1.1, 1.3, 3.1, 17.1-5, 18.1-3, 19.1-3, 22.1 and 25.1-2 were produced on MINITAB and the remaining plots were generated with R.

Bibliography

Ahovuo-Saloranta, A., Hiiri, A., Nordblad, A., Worthington, H., Makela, M., (2004). Pit and fissure sealants for preventing dental decay in the permanent teeth of children and adolescents (Cochran Review). In *The Cochran Library*, Issue 3. John Wiley & Sons, Ltd, Chichester, UK.

Albert, X., Pereiró, I., Sanfelix, J., Gosalbes, V., and Perrota, C. (2004). Antibiotics for preventing recurrent urinary tract infection in non-pregnant women (Cochran Review). In *The Cochran Library*, Issue 3. John Wiley & Sons, Ltd, Chichester, UK.

Altman, D.G. (1991) *Practical Statistics for Medical Research*. Chapman and Hall, London.

Amess, J.A.L., Burman, J.F., Rees, G.M., Nacekievill, D.G., and Mollin, D.L. (1978). Megaloblastic haemopoiesis in patients receiving nitrous oxide. *Lancet*, (2):339–342.

Anscombe, F. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika*, 35:266–254.

Aspin, A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika*, 35:88–96.

Azorin, P.F. (1953). Sobre la distribución t no central I,II, *Trabajos de Estadistica*, 4:173–198, 307–337.

Bar-Lev, S.K. and Enis, P. (1988). On the classical choice of variance stabilizing transformations and an application for a Poisson variate. *Biometrika*, 75(4):803–804.

Berger, J. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? [With discussion.] *Statistical Science*, 18(1):1–32.

Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence. [With discussion.] *Journal of the American Statistical Association*, 82:112–139.

Berger, J., Boukai, B. and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science*, 12:133–160

Bhattacharya, B. and Habtzghi, D. (2002). Median of the p Value under the alternative hypothesis. *The American Statistician*, 56(3):202–206.

Bickel, P. and Doksum, K. (1990). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco.

Biggerstaff, B. and Tweedie, R. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, 16:753–768.

Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence Elena Kulinskaya, Stephen Morgenthaler, and Robert G. Staudte

© 2008 John Wiley & Sons, Ltd

Brown, B.W. and Hollander, M. (1977). *Statistics: A Biomedical Introduction*. John Wiley & Sons, Inc., New York.

Brown, L. and Li, X. (2005). Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference,* 130:359–375.

Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. [With discussion.] *Journal of the American Statistical Association,* 82:106–111, 123–135.

Chernoff, H. and Lehmann, E. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statistics,* 25(3):579–586.

Cochran, W. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society,* 4(Suppl.):102–118.

Cochran, W. (1954). The combination of estimates from different experiments. *Biometrics,* 10(1):101–129.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd edition. Lawrence Earlbaum Associates, Hillsdale, NJ.

Colditz, G., Brewer, T., Berkey, C.S., Wilson, M.E., Burdick, E., Fineberg, H. V., and Mosteller, F. (1994). Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. *Journal of the American Medical Assiciation,* 271:698–702.

Collins, R., Yusuf, S., and Peto, R. (1985). Overview of randomised trials of diuretics in pregnancy. *British Medical Journal,* 291:97–104.

D'Agostino, R. and Stephens, M., editors (1986). *Goodness-of-fit Techniques*. Marcel Dekker, New York.

Decker, R. and Fitzgibbon, D. (1991). The normal and Poisson approximations to the binomial: a closer look. Technical report no. 82.3, Department of Mathematics, Univeristy of Hartford.

Dempster, A. and Schatzoff, M. (1965). Expected significance level as a sensitivity index for test statistics. *Journal of the American Statistical Association,* 60:420–436.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials,* 7:177-188.

Devleeschouwer, N., Libeer, J., Martens, F., Neels, H., Damme, V.M., Verstraete, A., Deveaux, M., and Wallemacq, P. (2004). Blood alcohol testing: comparison of the performance obtained with the different methods used in the Belgian external quality assessment schemes. *Clinical Chemistry Laboratory Medicine,* 42(1):57–61.

Diggle, P. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

Donahue, R. (1999). A note on information seldom reported via the p-value. *The American Statistician,* 53(4):303–306.

Efron, B. (1982). Transformation Theory: how normal is a family of distributions? *The Annals of Statistics,* 10(2):323–339.

Fisher, R., (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture for Great Britain,* 33:503–513.

Fisher, R. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Givens, G.H., Smith, D.D., and Tweedie, R.L. (1997). Publication in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science,* 12(4):221–250.

Glass, G. (1976). Primary, secondary and meta-analysis of research. Educational Researcher, 5:3–8.

Goodman, S. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine,* 11:875–879.

Goodman, S. (1998). P Value, in *Encyclopedia of Biostatistics*, Vol. 4 (ed. Armitage, P. and Colton, T.). John Wiley & Sons, Ltd, London, pp. 3233–3237.

Greenwood, P. and Nikulin, M. (1996). *A Guide to Chi-Squared Testing*. John Wiley & Sons, Inc., New York.

Hardy, R. and Thompson, S. (1996). A likelihood approach to meta-analysis. *Statistics in Medicine*, 15:619–629.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-analysis*. Academic Press, Orlando, FL.

Hoenig, J. and Heisey, D. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55:19–24.

Holland, P. (1973). Covariance stabilizing transformations. *Annals of Mathematical Statistics*, 1:84–92.

Hubbard, R. and Bayarri, M. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. [With discussion]. *The American Statistician*, 57(3):171–182.

Huffman, M. (1984). An improved approximate two-sample Poisson test. *Applied Statistics*, 33:224–226.

Hung, H., O'Neill, R., Bauer, R., and Kohne, K. (1997). The behaviour of the p value when the alternative is true. *Biometrics*, 53:11–22.

James, G.S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38:324–329.

Jané-Llopis, E., Hosman, C., Jenkins, R., and Anderson, P. (2003). Predictors of efficacy in depression prevention programmes. *British Journal of Psychiatry*, 183:384–397.

Johnson, N., Kotz, S., and Kemp, A. (1993). *Univariate Discrete Distributions*, 2nd edition. John Wiley & Sons, Inc., New York.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, volume 2. John Wiley & Sons, Inc., New York.

Karlin, S. and Taylor, H. (1975). *A First Course in Stochastic Processes*. Academic Press, New York.

Kieler, H., Axelsson, O., Haglund, B., Nilsson, S., Salvesen, K.A. (1998). Routine ultrasound screening in pregnancy and the children's subsequent handedness. *Early Human Development*, 50:233–245.

Kulinskaya, E. and Staudte, R.G. (2007). Confidence intervals for the standardized effect arising in comparisons of two normal populations. *Statistics in Medicine*, 26(14):2853–2871.

Lachin, J. (2000). *Biostatistical Methods: The Assesment of Relative Risks*. Wiley Series in Probability & Statistics. John Wiley & Sons, Inc., New York.

Larsen, R.J. and Marx, M.L. (1986). *An Introduction to Mathematical Statistics and its Applications*, (2nd edition). Prentice-Hall, Englewood Cliffs, NJ.

Laubscher, N. (1960). Normalizing the noncentral t and F distributions. *The Annals of Mathematical Statistics*, 31(4):1105–1112.

Lehmann, E. (1986). *Testing Statistical Hypotheses*, 2nd edition. John Wiley & Sons, Inc., New York.

Manly, B. (1991). *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall, 1991, London.

Manocha, S., Choudhuri, G. and Tandon, B. (1986). A study of dietary intake in pre- and post-menstrual period. *Human Nutrition: Applied Nutrition*, 40A:213–216.

McCullagh, P. and Nelder, J.A. (1999). *Generalized Linear Models*, 2nd edition. Chapman and Hall, London.

Mulrow, C., Chiquette, E., Angel, L., Cornell, J., Summerbell, C., Anagnosetelis, B., Brand, M., Grimm, R. (204). Dieting to reduce body weight for controlling hypertension in adults (Cochran Review). In *The Cochran Library*, Issue 3. John Wiley & Sons, Ltd, Chichester, UK.

Mumford, E., Schlesinger, H., Glass, C., Patrie, G., and Cuerdon, T. (1984). A new look at evidence about reduced cost of medical utilization following mental health treatment. *American Journal of Psychiatry*, 141:1145–1158.

Nel, D., van der Merwe, C.A., and Moser, B. (1990). The exact distributions of the univariate and multivariate Behrens-Fisher statstics with a comparison of several solutions in the univariate case. *Communications in Statistics – Theory and Methods*, 19:279–298.

Neyman, J. (1950). *First Course in Probability and Statistics*. Henry Holt, New York.

Preece, D., Ross, G. J. S., and Kirby, P. (1988). Bortkewitsch's Horse-Kicks and the Generalised Linear Model. *The Statistician*, 37:313–318.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd edition. John Wiley & Sons, Inc., New York.

Royall, R. (1997). *Statistical Evidence*, Monographs on Statistics and Applied Probability, Vol. 71. Chapman & Hall, London.

Sackowitz, H. and Samuel-Cahn, E. (1999). P values as random variables, expected P values. *The American Statistician*, 53(4):326–331.

Salvesen, K. A., Vatten, L.J., Eik-Nes, S.H., Hugdahl, K., and Bakketeig, L.S. (1993). Routine ultrasonography in utero and subsequent handedness and neurological development. *British Medical Journal*, 307(6897):159–164.

Schervish, M.J. (1996). P values: what they are and what they are not. *The American Statistician*, 50:203–206.

Selke, T., Bayarri, M., and Berger, J. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55:62–71.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, Probability and Statistics. John Wiley & Sons, Inc., New York.

Sutton, A., Abrams, K., Jones, D., Sheldon, T., and Song, F. (2000). *Methods for Meta-Analysis in Medical Research*. John Wiley & Sons, Ltd, Chichester, UK.

Thompson, S. and Higgins, J. (2002). How should meta-regression analyses by undertaken and interpreted? *Statistics in Medicine*, 21:1559–1574.

Tweedie, R., Scott, D., Biggerstaff, B., and Mengersen, K. (1996). Bayesian meta-analysis, with application to studies of ETS and lung cancer. *Lung Cancer*, 14(Suppl. 1):S171–S194.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Watson, G. (1957). The χ^2 goodness-of-fit test for normal distributions. *Biometrika*, 44(3/4):336–348.

Welch, B. (1938). The significance of the difference between two means when the variances are unequal. *Biometrika*, 29:350–361.

Welch, B. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34:28–35.

Welch, B.L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38:330–336.

Ziv, G. and Sulman, F.G. (1972). Binding of antibiotics to bovine and ovine serum. *Antimicrobial Agents and Chemotherapy*, 2:206–213.

Index

R functions website, xiv

Ahovuo-Saloranta, A. et al., 8
Albert, X. et al., 54, 157
Altman, D., 27, 43
Amess, J. et al., 43
ANOVA, **195**
Anscombe, F., 58, 132, 139
Aspin, A., 177
Azorin, A., 161

Balakrishnan, N., 160, 183
Bar-Lev, S.K., 132
Bayarri, M.J., 121
Berger, J.O., 121
Berger, R.L., 121
Bhattacharya, B., 118
Bickel, P.J., 126
Biggerstaff, B.J., 56, 217, 226, 229
Boukai, B., 121
Brown, B.W., 20
Brown, L., 11, 149

calibration scale, 4–6, 117
  see also evidence, 115
  see also *p*-value, 121
Casella, G., 121
Chernoff, H., 73
Cochran Review, 157, 175
Cochran's *Q*, 8, 10, 77, 85, **210**, 212
Cochran, W.G., 209
Cohen's *d*, 177
Cohen, J., 177, 180
Colditz, G. et al., 67, 100, 101
Collins, B. et al., 56
combined evidence
  fixed effects, 4, 8, 10, **222**
  random effects, 8, 10, **226**
confidence bounds
  chi-squared non-centrality parameter, 188
confidence intervals
  after combining studies
    fixed effects model, 222, 223
    random effects model, 226
  chi-squared non-centrality parameter, 189
  F non-centrality parameter, 203
  one-sample binomial, 142
    by log-transformation, 146
    by Poisson approximation, 147
  one-sample normal
    standardized effect, 167
  one-sample Poisson, 132
  t non-centrality parameter, 168
  two-sample binomial
    risk difference, 155
    standardized effect, 153
  two-sample normal
    effect, 177
    standardized effect, 180
  two-sample Poisson, 67
  via Key Inferential Function, 128
correlation effect size, 12, 150

D'Agostino, R.B., 71
Decker, R.D., 147

Dempster, A., 118
DerSimonian, R., 225, 229, 232
Devleeschouwer, N. et al., 21
Diggle, P.J., 75
distribution
  non-central chi-squared, **183**, 196
  non-central *F*, **196**
  non-central *t*, **160**
  truncated normal, **235**
Doksum, K., 126
Donahue, R., 118

effect, 16, **114**, 159
  see also raw effect, 149
effect size, 12, 180, 243
  see also standardized effect, 12
Efron, B., 132, 138
Enis, P., 132
evidence, **113**, 165
  calibration scale for, 5, **15–19**, 113
  chi-squared statistic, **184**
  corrected, 165
  F statistic, **199**
  heterogeneity
    fixed effects, 77, **212**
    random effects, 80, **217**
    transformed effects, **213**
  one-sample binomial, **139**
  one-sample Poisson, **130**
  one-sided, **114**
    connection to *p*-values, 6, 115
    properties of, **115**
    weak, moderate and strong, 5, 115
  Student's *t*-statistic, **161**
  Student's *t*-test versus *Z*-test, 170
  two-sample binomial, **150**
  two-sample Poisson
    conditional, 66
    unconditional, 64
  two-sided, 134
    relation to one-sided, 19
    via chi-squared distribution, 135
    via doubling the *p*-value, 137
Examples
  bellbirds arriving to feed nestlings, 74
  blood alcohol testing, 21
  characteristics of antibiotics, 42
  counts of cells growing in culture, 129, 131
  daily energy intake compared to a fixed level, 27
  Darwin's data on *Zea mays*, 28
  deaths by horsekicks, 60, 81, 87
  depression prevention programs, 108
  diuretics in pregnancy and risk of pre-eclamsia, 56, 84
  effect of open education on student creativity, 100
  effect of psychotherapy on hospital length of stay, 37, 83
  environmental tobacco smoke, 107
  filling containers, 19
  glass ionomer versus resin sealants, 9
  recurrent urinary tract infections, 50, 54, 157
  red cell folate levels, 43
  stability of blood samples, 20
  systolic blood pressure, 35, 82, 88, 159, 168, 175, 181, 229
  ultrasound and left-handedness, 49
  vaccination for prevention of tuberculosis, 67, 101

Fisher, R.A., 28, 115
Fitzgibbon, D.J., 147
fixed effects model, 8, **195**, 206
funnel plot, **233**

Givens, G.H., 240
Glass, G.V., 177
Goodman, S., 113, 119
goodness-of-fit testing, 74
Greenwood, P., 72, 73

Habtzghi, D., 118
Hardy, R., 56

Hedges, L.V., 100, 177, 180
Heisey, D.M., 123
heterogeneity of effects
  see also evidence, 209
  standardized, 8, 231
  transformed, 8, 10, 221
Higgins, J., 95
Hoenig, J.M., 123
Holland, P.W., 241
Hollander, M., 20
homogeneity of effects, xiv, 8, 77, 85, 131, 209, 211, 212, 214, 216
Hubbard, R., 121
Huffman, M.D., 64
Hung, H. et al., 118
hyperbolic cosine function, **196**
hyperbolic sine function, **161**

James, G.S., 209
Jané-Llopis, E., 108
Johnson, N.L., 126, 147, 160, 183

Karlin, S., 74
Kemp, A.W., 126, 147
Key Inferential Function, 6, **127**, 222
  chi-squared distribution, **185**
  F distribution, **203**
  one-sample binomial, 7, **140**
  one-sample Poisson, **130**
  Student's *t* distribution, 7, **165**
  two-sample binomial, **150**
  two-sample normal
    pooled *t*, **179**
    Welch *t*, **178**
  two-sample Poisson
    conditional, **66**
    unconditional, **64**
Kotz, S., 126, 147, 160, 183
Kulinskaya, E., 178, 180, 209, 216, 224

Lachin, J., 155
Laird, N., 225, 229, 232
Laubscher, N.F., 162, 196, 197
Lehmann, E.L., 66, 73
Li, X., 12, 149

Manly, B., 28
Manocha, S. et al., 27
McCullagh, P., 96
meta-regression, **95–100**
Moser, B.K., 177
Mulrow, C.D. et al., 82, 88, 159, 175
Mumford, E. et al., 36, 83

Nel, D.G., 177
Nelder, J.A., 96
Neyman, J., 123
Nikulin, M., 72, 73
nuisance parameters, 87, 223

odds ratio, 54, 145, **149**, 157
Olkin, I., 100, 177, 180

p-value, 113
  as conditional probability, 116
  Bayesian calibration, 121
  random, **117**
    conditonal evidence in, 121
    confidence interval for median, 120
    expected *p*-value, 118
    interpretation of, 118
    properties of distribution, 118
Pitman efficacy, 243
power, 245
  via expected evidence, 17
power function
  via Key Inferential Function, 128
Preece, D., 60
probit, 6, 116, 117
prototypical model, xiii, 15, **114**, 117, 121
publication bias
  censoring model, 238
  truncation model, 121, 235

random effects model, **206**, 216
random transformed effects model, 8, **224**
Rao, C.R., 184

raw effect, 6, 7, 12, 28, 37, 52, 56, 81, 82, 88, 95, 175, 176, 180
  see also effect, 149
relative risk, 54, 99, 144, **149**, 155
risk difference, **149**

Sackowitz, H., 118
sample size calculations
  for evidence, 17, 127
    chi-squared, 190
    one sample Poisson, 132
    one-sample binomial, 48
    pooled *t*-test, 32
    Student's *t*-test, 163
    two-sample binomial, 151
    Welch *t*-test, 34, 179
  for power, 245
Samuel-Cahn, E., 118
Schatzoff, M., 118
Schervish, M.J., 113
Sellke, T., 121
Smith, D.D., 240
standardized effect, 12, 16, 32, **114**
  for one population, 12
  for two populations, 12
  one-sample normal, 159
  see also effect size, 12
  two-sample binomial, 150
  two-sample normal, 175, 177
statistical evidence
  see also evidence, 1
Staudte, R.G., 178, 180, 209, 216, 224
Stephens, M.A., 71
Sutton, A.J. et al., 67

Taylor, H.M., 74
Thompson, S., 56, 95
transformed effect
  see also effect, 221
Tweedie, R.L et al., 107
Tweedie, R.L., 54, 217, 226, 229, 240

van der Merwe, C.A., 178
van der Vaart, A. W., 243
variance component, 217, 229, 232
variance stabilization, **126**
variance stabilizing transformation, **126**
  asymptotic, 241
  Cochran's Q, **217**
  non-central chi-squared statistic, **184**
  non-central F statistic, **199**
  one-sample binomial, 7, **139**
  one-sample Poisson, **130**
  Student's *t*, 7, **161**
  two-sample binomial
    risk difference, **150**
  two-sample Poisson
    conditional, **66**
    unconditional, **64**
  Welch *t*, **178**
vst
  see variance stabilizing transformation, 125

Wang,Y., 121
Watson, G.S., 73
weighted mean, 183, 231
weights, 3, 7, 14, 97, 213, 217, 223, 225, 226
  estimated, 97, 102, 209, 213, 229
  known, 96, 209, 212, 213, 220
  unknown, 211, 212
Welch, B.L., 175, 177, 209

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors
David J. Balding, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J.B. Kadane, Geert Molenberghs, David W. Scott, Adrian F.M. Smith, Sanford Weisberg

Editors Emeriti
Vic Barnett, J. Stuart Hunter, David G. Kendall, Jozef L. Teugels

The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis
AGRESTI · Categorical Data Analysis, Second Edition
ALTMAN, GILL and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, Third Edition
*ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN and NAGARAJA · Records
*ARTHANARI and DODGE · Mathematical Programming in Statistics
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, Third Edition
BARNETT · Environmental Statistics: Methods & Applications
BARNETT and LEWIS · Outliers in Statistical Data, Third Edition
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
BELSLEY, KUH and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

*Now available in a lower priced paperback edition in the Wiley Classics Library.

BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, Third Edition
BERNARDO and SMITH · Bayesian Theory
BERRY, CHALONER and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BHAT and MILLER · Elements of Applied Stochastic Processes, Third Edition
BHATTACHARYA and JOHNSON · Statistical Concepts and Methods
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BIEMER, GROVES, LYBERG, MATHIOWETZ and SUDMAN · Measurement Errors in Surveys
BILLINGSLEY · Convergence of Probability Measures, Second Edition
BILLINGSLEY · Probability and Measure, Third Edition
BIRKES and DODGE · Alternative Methods of Regression
BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE and MURTHY · Reliability: Modeling, Prediction and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, Second Edition
BOLLEN · Structural Equations with Latent Variables
BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOSQ and BLANKE · Inference and Prediction in Large Dimensions
BOULEAU · Numerical Methods for Stochastic Processes
BOX · Bayesian Inference in Statistical Analysis
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building
BOX, HUNTER and HUNTER · Statistics for Experimenters: Design, Innovation and Discovery, Second Edition
BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BRUNNER, DOMHOF and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
BUCKLEW · Large Deviation Techniques in Decision, Simulation and Estimation
CAIROLI and DALANG · Sequential Stochastic Optimization
CASTILLO, HADI, BALAKRISHNAN and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
CHAN · Time Series: Applications to Finance
CHATTERJEE and HADI · Regression Analysis by Example, Fourth Edition
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, Third Edition
CHERNICK · Bootstrap Methods: A Practitioner's Guide
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÉS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, Second Edition
CLARKE and DISNEY · Probability and Random Processes: A First Course with
Applications, Second Edition
*COCHRAN and COX · Experimental Designs, Second Edition
CONGDON · Applied Bayesian Modelling
CONGDON · Bayesian Models for Categorical Data
CONGDON · Bayesian Statistical Modelling
CONGDON · Bayesian Statistical Modelling, Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

CONOVER · Practical Nonparametric Statistics, Second Edition
COOK · Regression Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
CORNELL · Experiments with Mixtures, Designs, Models and the Analysis of Mixture Data, Third Edition
COVER and THOMAS · Elements of Information Theory
COX · A Handbook of Introductory Statistical Methods
*COX · Planning of Experiments
CRESSIE · Statistics for Spatial Data, Revised Edition
CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, Sixth Edition
*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, Second Edition
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, Third Edition
*DEGROOT, FIENBERG and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications
DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, Second Edition
*DOOB · Stochastic Processes
DOWDY, WEARDEN and CHILKO · Statistics for Research, Third Edition
DRAPER and SMITH · Applied Regression Analysis, Third Edition
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, Second Edition
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, Third Edition
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
EDLER and KITSOS (editors) · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS and PEACOCK · Statistical Distribution, Third Edition
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, Third Edition, Revised; Volume II, Second Edition
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
FITZMAURICE, LAIRD and WARE · Applied Longitudinal Analysis
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, Second Edition
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, Second Edition
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GEISSER · Modes of Parametric Statistical Inference
GELMAN and MENG (editors) · Applied Bayesian Modeling and Casual Inference from Incomplete-data Perspectives

*Now available in a lower priced paperback edition in the Wiley Classics Library

GEWEKE · Contemporary Bayesian Econometrics and Statistics

GHOSH, MUKHOPADHYAY and SEN · Sequential Estimation

GIESBRECHT and GUMPERTZ · Planning, Construction and Statistical Analysis of Comparative Experiments

GIFI · Nonlinear Multivariate Analysis

GIVENS and HOETING · Computational Statistics

GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems

GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, Second Edition

GOLDSTEIN and LEWIS · Assessment: Problems, Development and Statistical Issues

GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

GROSS and HARRIS · Fundamentals of Queueing Theory, Third Edition

*HAHN and SHAPIRO · Statistical Models in Engineering

HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners

HALD · A History of Probability and Statistics and their Applications Before 1750

HALD · A History of Mathematical Statistics from 1750 to 1930

HAMPEL · Robust Statistics: The Approach Based on Influence Functions

HANNAN and DEISTLER · The Statistical Theory of Linear Systems

HEIBERGER · Computation for the Analysis of Designed Experiments

HEDAYAT and SINHA · Design and Inference in Finite Population Sampling

HEDEKER and GIBBONS · Longitudinal Data Analysis

HELLER · MACSYMA for Statisticians

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design

HINKELMANN and KEMPTHORNE · Design and analysis of experiments, Volume 2: Advanced Experimental Design

HOAGLIN, MOSTELLER and TUKEY · Exploratory Approach to Analysis of Variance

HOAGLIN, MOSTELLER and TUKEY · Exploring Data Tables, Trends and Shapes

*HOAGLIN, MOSTELLER and TUKEY · Understanding Robust and Exploratory Data Analysis

HOCHBERG and TAMHANE · Multiple Comparison Procedures

HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, Second Edition

HOEL · Introduction to Mathematical Statistics, Fifth Edition

HOGG and KLUGMAN · Loss Distributions

HOLLANDER and WOLFE · Nonparametric Statistical Methods, Second Edition

HOSMER and LEMESHOW · Applied Logistic Regression, Second Edition

HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data

HUBER · Robust Statistics

HUBERTY · Applied Discriminant Analysis

HUNT and KENNEDY · Financial Derivatives in Theory and Practice, Revised Edition

HUSKOVA, BERAN and DUPAC · Collected Works of Jaroslav Hajek—with Commentary

HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data

IMAN and CONOVER · A Modern Approach to Statistics

JACKSON · A User's Guide to Principle Components

JOHN · Statistical Methods in Engineering and Quality Assurance

JOHNSON · Multivariate Statistical Simulation

JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz

JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, Fifth Edition

JUDGE, GRIFFITHS, HILL, LU TKEPOHL and LEE · The Theory and Practice of Econometrics, Second Edition

JOHNSON and KOTZ · Distributions in Statistics

JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present

JOHNSON, KOTZ and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

JOHNSON, KOTZ and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, Second Edition
JOHNSON, KOTZ and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ and KEMP · Univariate Discrete Distributions, Second Edition
JUREČKOVÁ and SEN · Robust Statistical Procedures: Asymptotics and Interrelations
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE and SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, Second Edition
KARIYA and KURATA · Generalized Least Squares
KASS and VOS · Geometrical Foundations of Asymptotic Inference
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, Second Edition
KHURI, MATHEW and SINHA · Statistical Tests for Mixed Linear Models
*KISH · Statistical Design for Research
KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
KLUGMAN, PANJER and WILLMOT · Loss Models: From Data to Decisions
KLUGMAN, PANJER and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions
KOTZ, BALAKRISHNAN and JOHNSON · Continuous Multivariate Distributions, Volume 1, Second Edition
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, Second Edition
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST and GREENHOUSE · Case Studies in Biometry
LARSON · Introduction to Probability Theory and Statistical Inference, Third Edition
LAWLESS · Statistical Models and Methods for Lifetime Data, Second Edition
LAWSON · Statistical Methods in Spatial Epidemiology, Second Edition
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
LEE and WANG · Statistical Methods for Survival Data Analysis, Third Edition
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LINDVALL · Lectures on the Coupling Method
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data, Second Edition
LLOYD · The Statistical Analysis of Categorical Data
LOWEN and TEICH · Fractal-Based Point Processes
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MALLOWS · Design, Data and Analysis by Some Friends of Cuthbert Daniel

*Now available in a lower priced paperback edition in the Wiley - Interscience Paperback Series.

MANN, SCHAFER and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
MANTON, WOODBURY and TOLLEY · Statistical Applications Using Fuzzy Sets
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and practice
MARDIA and JUPP · Directional Statistics
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods
MASON, GUNST and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, Second Edition
MCCULLOCH and SERLE · Generalized, Linear and Mixed Models
MCFADDEN · Management of Data in Clinical Trials
MCLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
MCLACHLAN, DO and AMBROISE · Analyzing Microarray Gene Expression Data
MCLACHLAN and KRISHNAN · The EM Algorithm and Extensions
MCLACHLAN and PEEL · Finite Mixture Models
MCNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MICKEY, DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, Third Edition
*MILLER · Survival Analysis, Second Edition
MONTGOMERY, PECK and VINING · Introduction to Linear Regression Analysis, Fourth Edition
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
MULLER and STEWART · Linear Model Theory: Univariate, Multivariate and Mixed Models
MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis and Nonlinear Optimization
MURTHY, XIE and JIANG · Weibull Models
MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Second Edition
MYERS, MONTGOMERY and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
**NELSON · Accelerated Testing, Statistical Models, Test Plans and Data Analysis
**NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
OKABE, BOOTS, SUGIHARA and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, Second Edition
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PALTA · Quantitative Methods in Population Health: Extentions of Ordinary Regression
PANJER · Operational Risks: Modeling Analytics
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
*PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO and TSAY · A Course in Time Series Analysis
PIANTADOSI · Clinical Trials: A Methodologic Perspective
PORT · Theoretical Probability for Applications
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
PRESS · Bayesian Statistics: Principles, Models and Applications
PRESS · Subjective and Objective Bayesian Statistics, Second Edition
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach

*Now available in a lower priced paperback edition in the Wiley Classics Library.
**Now available in a lower priced paperback edition in the Wiley - Interscience Paperback Series.

PUKELSHEIM · Optimal Experimental Design
PURI, VILAPLANA and WERTZ · New Perspectives in Theoretical and Applied Statistics
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
QIU · Image Processing and Jump Regression Analysis
RAO · Linear Statistical Inference and its Applications, Second Edition
RAUSAND and HÃ˜YLAND · System Reliability Theory: Models, Statistical Methods and Applications, Second Edition
RENCHER · Linear Models in Statistics
RENCHER · Methods of Multivariate Analysis, Second Edition
RENCHER · Multivariate Statistical Inference with Applications
RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROBINSON · Practical Strategies for Experimenting
ROHATGI and SALEH · An Introduction to Probability and Statistics, Second Edition
ROLSKI, SCHMIDLI, SCHMIDT and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROSSI, ALLENBY and MCCULLOCH · Bayesian Statistics and Marketing
ROUSSEEUW and LEROY · Robust Regression and Outline Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement, Second Edition
SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN and SCOTT (editors) · Sensitivity Analysis
*SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation and Application
SCHOTT · Matrix Analysis for Statistics
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice and Visualization
*SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · Multivariate Observations
SEBER and LEE · Linear Regression Analysis, Second Edition
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: Its Only a Game!
SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order and Shape Restrictions
SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL and MECKE · Stochastic Geometry and Its Applications, Second Edition
STOYAN and STOYAN · Fractals, Random and Point Fields: Methods of Geometrical Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

STYAN · The Collected Papers of T. W. Anderson: 1943–1985 SUTTON, ABRAMS, JONES, SHELDON and SONG · Methods for Meta-Analysis in Medical Research
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, Second Edition
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÃ‘A and STIGLER (editors) · Box on Quality and Discovery: with Design, Control and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VAN BELLE, FISHER, HEAGERTY and LUMLEY · Biostatistics: A Methodology for the Health Sciences, Second Edition
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEERAHANDI · Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models
WEISBERG · Applied Linear Regression, Second Edition
WELISH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, Second Edition
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, Second Edition
WU and HAMADA · Experiments: Planning, Analysis and Parameter Design Optimization
WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches
YANG · The Construction Theory of Denumerable Markov Processes
YOUNG, VALERO-MORA and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics
*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZELTERMAN · Discrete Distributions: Applications in the Health Sciences
ZHOU, OBUCHOWSKI and McCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.